MULTIMODAL E-LEARNING:
AN EMPIRICAL STUDY

Musa Khalifa A Faneer

PhD

Department of Computer Sciences and Media Technology
Faculty of Technology
De Montfort University

2015

MULTIMODAL E-LEARNING:
AN EMPIRICAL STUDY

An experimental study to investigate the usability, user experience and affective state aspects of using multimodal communication metaphors in the presentation of learning contents of e-learning interfaces and the production of empirically derived guidelines for the employment of such metaphors in user interfaces and software engineering process

Musa Khalifa A Faneer

Thesis Submitted for the Degree of Doctor of Philosophy in

Computer Science

Supervised by: Professor Dimitrios I. Rigas

Department of Media Technology, Faculty of Technology
De Montfort University
2015

Dedications

This thesis is dedicated to my family for their patience, support and constant love.

Musa Faneer

ABSTRACT

This empirical work aims to investigate the impact of using multimodal communication metaphors on e-learning systems' usability, overall user experience and affective state. The study proposed a triple evaluation approach to avoid the problem of conventional assessment relying only on usability measurements of efficiency, effectiveness and user satisfactions. Usability in that sense refers only to the functionality and pragmatic side of the product and neglects other aspects of the system. Learning is a cognitive and repetitive task, requiring learners' attention as well as their interest. Therefore, when delivering content, in addition to the pragmatic functionality, an e-learning system should provide a constructive overall user experience and positive affective state. Doing so will ensure user engagement, facilitate the learning process and increase learners' performance. The impact of using five different communication metaphors was evaluated in three dimensions using the proposed approach. Within the usability dimension, the evaluation criteria involved measuring system efficiency, effectiveness, user satisfaction and learning performance. Within the user experience dimension, the evaluation criteria involved measuring pragmatic aspects of the user experience, the hedonic aspects of user experience in terms of stimulation as well as identification and the overall system attractiveness. Within the affective state dimension a self-assessments manikin technique was used in conjunction with biofeedback measurements, and users' valence, arousal and dominance were measured. The study found that system attractiveness and the hedonic user experience had a profound impact on users' learning performance and attitude toward the tested system. Furthermore, they influenced users' views and judgement of the system and its usability. The communication metaphors were not equal in their influence within the evaluation criteria. Empirically derived guidelines were produced for the use and integration of these metaphors in e-learning systems. The outcome of the study highlights the need to use the triple evaluation approach in the assessment of e-learning interfaces prior to their release for better adoption and acceptance by end users.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

3

3E: Expressing Experiences and Emotions.

A

ANS: Autonomic Nervous System.

ATT: Attractiveness·

C

CBT: Computer Based Training.

CNS: Central Nervous System.

CD-ROM: Compact Disc Read-Only Memory.

CTML: Cognitive Theory of Multimedia Learning.

D

DVD: Digital Versatile Disc.

E

EEG: Electroencephalography.

EDA: Electro-dermal Activities.

G

GEW: Geneva Emotion Wheel.

GSR: Galvanic Skin Response.

GUI: Graphical User Interface.

H

HQ: Hedonic quality.

HQ - I: Hedonic Quality - Identity.

HQ - S: Hedonic Quality - Stimulation.

HRV: Heart Rate Variability.

I

ICT: Information and Communication Technologies (ICT).

IIS: Innovative Interactive Systems.

P

PSNS: Para-Sympathetic Nervous System.

PQ: Pragmatic Quality.

PAD: Pleasure, Arousal and Dominance.

PNS: Peripheral Nervous System.

S

SC: Skin Conductance.

SKT: Skin Temperature.

SPSS: Statistical Package for the Social Sciences.

SUS: System Usability Scale.

SNS: Sympathetic Nervous System.

SoNS: Somatic Nervous System.

U

UX: User Experience.

ACKNOWLEDGEMENT

I would like to express my deep and sincere gratitude to my supervisor **Professor Dimitrios I. Rigas** for his commitment, continuous support, patience, kindness, encouragement and invaluable guidance.

Also would like to thank Dr Amar Alaswed from the Libya cultural division for his help and support. Besides that would like to thank Mr Mike Cummings from the Affectiva team for facilitating the acquisition of the QSensor , Miss Silja Donadel from AttrakDiff support team for facilitating my work and allowing me to exceed the standard number of users allowed in their user experience test.

Last, but not least would like to thank the members of the IIS research group for their valuable comments and constructive discussions.

CHAPTER 1

1 THESIS OUTLINE

1.1 INTRODUCTION

Because computers affect all aspects of people's everyday lives, users expect more from them than just ease of use or achieving a particular task; they expect enjoyment and entertainment. Previous empirical studies [1, 2, 3 and 4] conducted by the Innovative Interactive Systems (IIS) research group have used multimodal metaphors such as audio, video, earcons, auditory icons, synthesized as well as recorded speech and avatars in e-learning interfaces. Like most new applications, the experimental platforms were tested for usability, as this has traditionally been perceived as an indicator for system quality [5]. Dillon [6] argues that it is clear that usability on its own, at least as it is currently understood, is inadequate for ensuring high quality user involvement with a new technology. Battarbee *et al*. [7] also agree that usability is important but that it is not sufficient to ensure the success of the system with end users. While usability can help people to take advantage of a system's functionality, usability also needs to open the way for fun and more pleasurable experiences. It is crucial to expand the classic usability evaluation techniques to include a more holistic set of user experience measures, beyond simple task evaluation of efficiency, effectiveness, and satisfaction. More attention must be given to user selections, preference, aesthetic feeling, frustration and sense of accomplishment.

User Experience (UX) is a new term used for system evaluation [8]. The term UX was initially introduced by Norman [9] and evolved to go beyond the traditional understanding of effectiveness, efficiency and satisfaction. Dillon [6] includes three key elements for successful user interaction with the system:

- Process, paying attention to what the user activities are and how they accomplish particular tasks provides clearer understanding of user preferences, navigations and system drawbacks.
- Outcomes, understanding user accomplishment goals gives an understanding of what the system does to help the user attain those goals.

- Affect, understanding user interaction, emotional reaction and feeling toward the system goes beyond the short-term satisfaction concept of usability measurements.

The International Standards Organization (ISO) defines user experience as *“a person's perception and responses that result from the use or anticipated use of a product, system or service”* [10, 11 and 12] proposes extending the definition of user satisfaction to include the following elements:

- Likability, how satisfied the user is with their experience of operating the system and accomplishments of pragmatic goals.
- Pleasure, how satisfied the user is with their achievement of accomplishing hedonic goals of motivating, identification and elicitation and accompanying emotional reactions [13].
- Comfort, to what degree the user is pleased and comfortable while they are using and interacting with the system.
- Trust, to what extent the user is satisfied and confident that the system will behave as intended.

According to Hassenzahl [14] UX differs greatly from usability, because of its emphasis on the following factors:

- The positive aspects of the user system relationship (enjoyment rather than frustration).
- The incorporation of hedonic (non-instrumental) quality aspects of the system.
- User experience focuses on the subjective side of system usage rather than the instrumental side and functionality of it, which is the main concern of usability evaluation (objective performance criteria).

Hassenzahl proposes the integration of user experience and usability in the user centred design. From the literature, we have found that usability on its own is not sufficient for the success of the system and more attention should be paid to user experience and emotion. A more holistic approach should be considered in order for the hedonic and pragmatic goals of the user to be met. A triple evaluation approach will be adapted, where users’ emotions, experiences, and usability assessment will be examined in order to deliver e-learning solutions which are emotionally affective, provide a positive experience and are subjectively and objectively desirable (Figure 1-1).

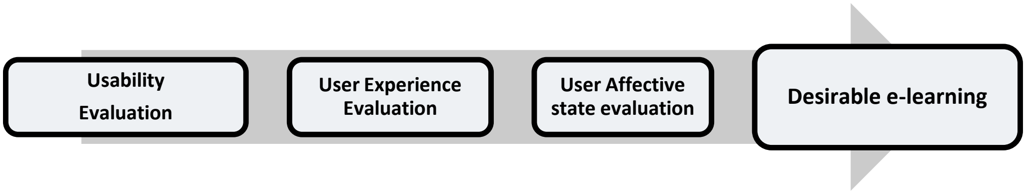


Figure 1-1: The proposed triple evaluation approach.

This thesis empirically investigates the roles of multimodal communication metaphors in delivering and promoting an optimized e-learning user experience. The experiment carried out is aimed at investigating the impact of using multimodal metaphors such as auditory icons, facially-expressive speaking avatars, earcons, classical music, and recorded speech, non-multimodal “visual only” in multimodal e-learning interfaces and their impact on:

- Interface usability in terms of efficiency, effectiveness, user satisfaction and learning performance.
- The overall user experience in terms of the system provided pragmatic as well as hedonic qualities and attractiveness.
- The user affective state in terms of arousal, valence and dominance.

This study is expected to aid in answering the following questions:

- How effective, efficient and satisfying is each communication metaphor?
- How affective is each metaphor?
- How hedonic, pragmatic and attractive is each metaphor?

1.2 AIMS AND OBJECTIVES

This research aims to investigate the role of multimodal communication metaphors in delivering an optimized e-learning experience in addition to producing empirically-derived guidelines for using and integrating various communication metaphors in e-learning interfaces. Multimodal communication metaphors investigated in this research are, auditory icons, earcons, facially-expressive avatars, classical music and recorded speech. The above-mentioned metaphors were tested in two scenarios:

- A multimodal scenario, where all communication metaphors were merged together and used in delivering the learning context.

- A unimodal scenario, where five conditions were created, with each condition dominated by a single communication metaphor and used to deliver a selected learning topic.
- The aim of this investigation was to evaluate the impact of utilizing such communication metaphors on the overall e-learning user experience concerning, system efficiency, effectiveness, hedonic quality, pragmatic quality, attractiveness, and user satisfaction as well as users' learning performance and affective state.

To fulfil the stated research aims two experimental studies were designed and executed. In the first study, the overall e-learning user experiences of two e-learning conditions were examined and evaluated. The first condition was controlled and referred to as a non-multimodal condition (NMMC). This condition included only text and graphics (visual only). The second condition was an experimental condition and was referred to as a multimodal condition (MMC). The multimodal condition employed the following communication metaphors "auditory icons", "facially expressive avatars", "earcons", "classical music", "recorded speech", in addition to texts and graphics. Both conditions were used to deliver similar learning topics with matching length and level of complexity. Within-subjects testing technique was adopted for the whole study. Each user tested both conditions and after going through each condition users were requested to perform recall and recognition tasks with comparable complexity levels (easy, moderate and difficult). The second experimental study separated the metaphors used by the multimodal platform in the previous study. The aim was to determine the impact that each metaphor had on the overall perceived enhancements. The study was conducted using a new platform with six conditions, presenting related learning topics with similar length and degree of complexity; each condition was designed and built to be dominated by only one communication metaphor and each condition was named after its dominating metaphor.

1.3 OVERALL HYPOTHESIS

Multimodal communication approach in e-learning platforms will enhance the overall e-learning user experience in comparison with non-multimodal approaches. Furthermore evaluating multimodal metaphors using a combined evaluation approach to assess the effectiveness, efficiency, satisfaction, hedonic quality, pragmatic quality, attractiveness, valence, arousal, and dominance will result in a better understanding of how to utilize such metaphors in order to optimize the overall e-learning user experience. The adapted research method involved a literature review and two experimental studies with subjective and objective assessment techniques; the techniques included using questionnaires, observations and biofeedback measurements. The data related to usability measurements (effectiveness, efficiency, user satisfaction and leaning) was collected using observation and questionnaires. The data related to user experience (hedonic quality, pragmatic quality and attractiveness) was obtained using a subjective technique utilizing AttrakDiff instrument [15]. Affective state data was collected using subjective as well as objective techniques including SAM scale [16] (for measuring valence, arousal, dominance) and biofeedback device (using QSensor from Affectiva) [17] (for measuring changes in skin conductance and temperature).

1.4 RESEARCH METHOD

Within the scope of this empirical research, the users were undergraduate and postgraduate university students from various schools. All of them were first-time users of the tested platforms and the data collected in each experimental session was analysed and discussed (it is crucial to mention here that the results obtained were content dependent on the material presented to the users) and finally a main conclusion was formulated. Accordingly, a set of empirically-derived guidelines and recommendations were produced for evaluating and using multimodal communication metaphors within e-learning interfaces. These guidelines are aimed at helping practitioners, designers and developers to deliver an optimized e-learning user experience. The proposed investigation model is illustrated in Figure 1-2 and the research method is delineated in the following subsections.

1.4.1 Literature Survey

The initial phase was to review the literature related to the research area, including the history of e-learning, e-learning systems, pedagogical principles, multimodal interaction, usability principles, user experience principles, user affective state and biofeedback (Chapter 2).

1.4.2 First Experiment

The first experiment was conducted to evaluate a non-multimodal e-learning condition, in comparison with a multimodal e-learning condition, within three areas including usability measurements (effectiveness, efficiency, user satisfaction and learning performance), user experience measurements (pragmatic quality, hedonic quality and attractiveness) and user affective state (valence, arousal, dominance, changes in skin conductance and temperature). The outcome of this experiment was the basis for the next experimental phase (see Chapter 3).

1.4.3 Second Experiment

This experiment was designed to separate the communication metaphors used in the first experimental phase and investigate the impact of each metaphor on the enhancements perceived in the first experiment. The evaluation criteria included usability measurements (effectiveness, efficiency, user satisfaction and learning performance), user experience measurements (pragmatic quality, hedonic quality and attractiveness) and users' affective state (valence, arousal, dominance, changes in skin conductance and temperature). This experiment can be found in Chapters 4, 5 and 6.

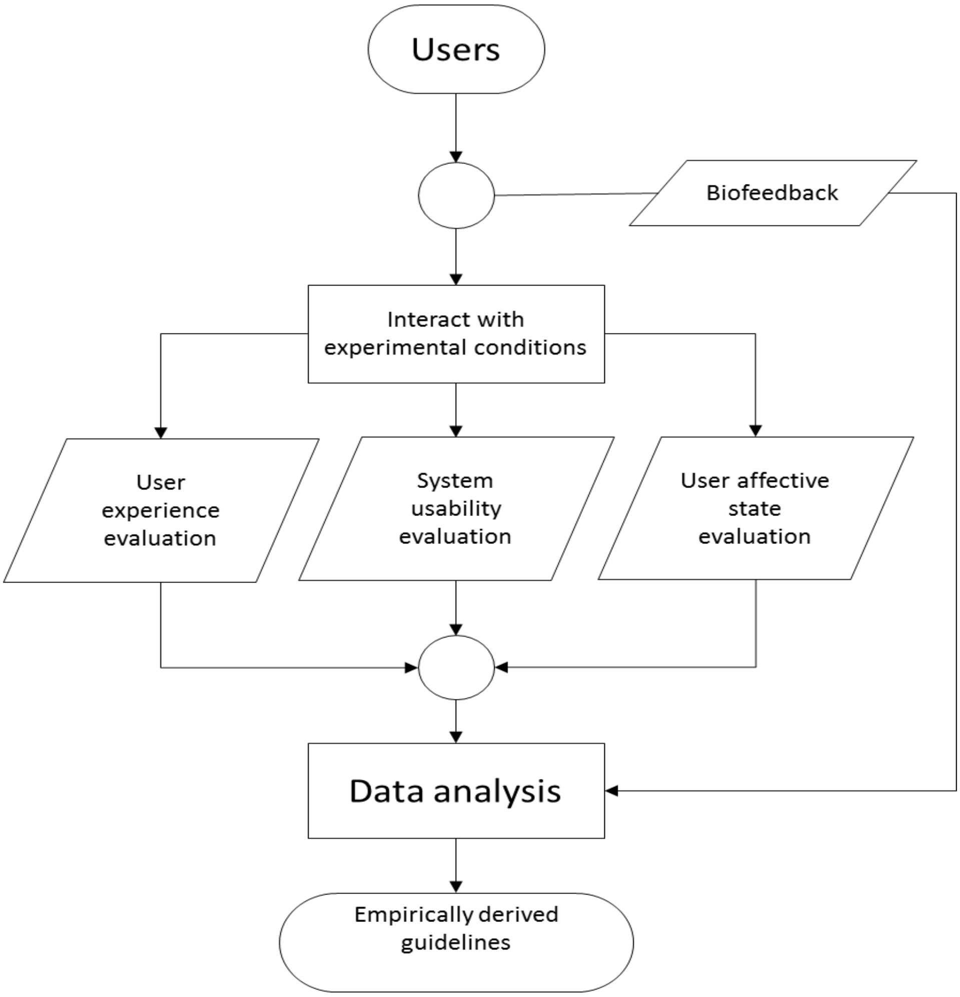


Figure 1-2: The proposed investigation model.

1.4.4 Conclusions, Empirically-Derived Guiding and Recommendations

This is the final chapter in which the findings are discussed, recommendations and empirically-derived guiding principles for integrating multimodal metaphors in e-learning interfaces are produced. These guidelines are expected to help in providing and optimizing e-learning user experience (Chapter 7).

1.5 THESIS CONTRIBUTION

These research findings contribute to the literature in the fields of e-learning, multimodal and human computer interaction (HCI). The contributions of this thesis are summarized in the following points:

- It presents a novel evaluation technique via a “triple evaluation approach” for assessing communication metaphors and multimodal user interfaces. This approach involves usability, user experience and affective state measurements.
- It provides empirically derived guidelines and recommendations for employing communication metaphors in e-learning and multimodal interaction interfaces.
- The presented evaluation approach used here could help e-learning and multimodal application developers to deliver an optimized and positively affective user experience.

1.6 THESIS STRUCTURE

The thesis is divided into seven chapters and a number of appendices. The subsequent sections refer to them (see Figure 1-3 for a quick summary).

1.6.1 Chapter 1: Introduction

This section provides an insight into the entire thesis and highlights the reviewed literature as well as the research hypothesis. It also describes the research methodology, the experimental work performed in addition to results and conclusion.

1.6.2 Chapter 2: Literature Review

This chapter offers an insight into the subject of e-learning, previous work, systems, concepts and pedagogical practices, multimodal metaphors, human computer interaction (HCI), usability evaluation, user experience and affective state.

1.6.3 Chapter 3: Experimental Phase I

This chapter describes, in detail, the empirical work performed in the first experiment. The study employed a within-subjects design to evaluate an e-learning platform with two conditions; a non-multimodal (control condition) was compared to a multimodal condition (experimental condition). The experiment involved 30 users who tested both conditions. The aim of the experiment was to evaluate the tested conditions with regards to usability measurements (effectiveness, efficiency as well as user satisfaction and learning performance), user experience measurements (pragmatic quality, hedonic quality and attractiveness) and users' affective state (self-assessment and changes in users' skin conductivity and temperature) while interacting with the e-learning platforms. The experiment paved the way for the next experimental phase.

1.6.4 Chapters 4, 5 and 6: Experimental Phase II

This experimental phase involved an e-learning platform with six conditions (including five experimental conditions and one controlled). This phase involved a group of 33 new user (they were randomly selected). The study used a within subjects design where each user tested all the experimented conditions. The aim of this empirical work was to investigate the impact each communication metaphor had on the perceived enhancements within the first experimental phase; the assessment criterion includes:

- Usability evaluation (in terms of effectiveness, efficiency user satisfaction and learning performance).
- User experience evaluation (in terms of pragmatic as well as hedonic qualities and attractiveness).
- User affective state evaluation (in terms of their emotional valence, arousal, dominance as well as skin conductance and temperature).
- This experiment is expected to reveal the role each communication metaphor could play in delivering an optimized e-learning experience.

1.6.5 Chapter 7: Findings, Empirically-Derived Guidelines and Recommendations

This chapter provides a summary of the empirical investigation performed in this research, research findings along with empirically derived guidelines and recommendations in addition to the limitations and proposed future work to extend this research.

1.6.6 Appendices

The appendices are divided into three sections and presents pre. and post. experimental shared evaluation items in addition to users' inputs and the collected data for the experimental sessions in relation to usability as well as user experience and affective state measurements.

Thesis Structure and the Experimental Phases Performed in This Study

Figure 1-3: Thesis structure and the performed experimental phases.

CHAPTER 2

LITERATURE REVIEW: E-LEARNING AND MULTIMODAL COMMUNICATION METAPHORS

2.1 INTRODUCTION

This section offers an overview and analysis of the literature in relation to the scope of the research area, Figure 2-1 encapsulates the main points covered in this review.

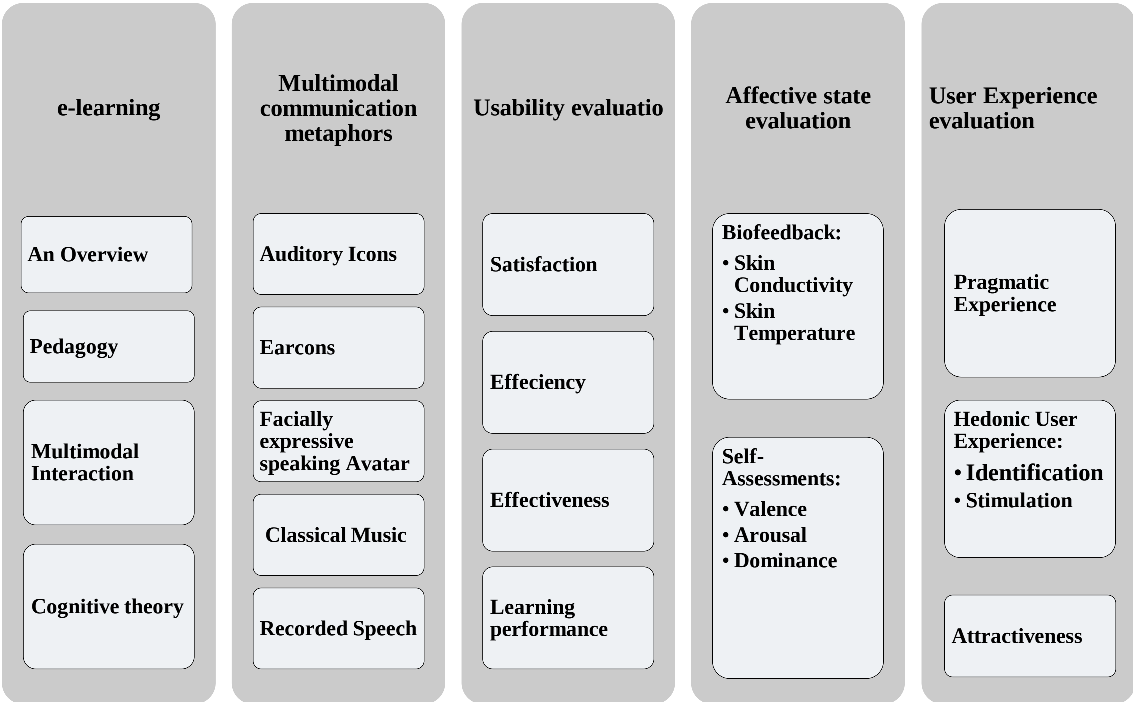


Figure 2-1: Main points of the literature review.

2.2 E-LEARNING

Increased computing power and communications' speed in addition to the extensive use of information and communication technologies (ICT) as well as the emergence of the World Wide Web, have transformed the way people conduct their daily activities; it has changed their way of buying goods, seeking medical advice, making flight reservations, finding friends or learning new skills. Furthermore, nowadays there are more individuals than ever before connected to the global network. Both public and private organizations are trying to capitalize on these technologies to reach their targeted audiences, not only to be more effective but also to survive in this knowledge-competitive era. One of the most prominent areas where ICT shows promising potential is the education sector, where there is a greater need for innovation along with creativity to meet the increasing demand for the creation, sharing and delivery of knowledge.

e-learning has fundamentally changed the way we think of education and learning and has evolved from being a mythical concept to being a main facility offered by schools and universities, as well as taking its place in future business plans for corporate training [18]. E-learning is emerging in different scenarios either as a standalone solution or as a complementary addition to lifelong learning and distance learning [19] and the e-learning market is growing gradually with an annual growth rate of 35.6% [20]. In recognition of the contribution of e-learning to greater knowledge retention, leading technology firms such as IBM and CISCO are building and developing a vast range of e-learning programs in both the public and private sectors. Advanced Distributed Learning [21] mentioned a couple of studies [22 and 23] indicating that, typically e-learning could save 30% of the time required to reach targeted instructional goals and increased students' skills. E-learning could save billions worldwide spent on conventional teaching methods and by freeing students from a classroom-restricted timetable and place, e-learning provides learners as well as instructors with the most flexible solution for interacting to acquire and share knowledge.

2.2.1 **What is E-Learning?**

According to Tavangarian *et al.* [24] e-learning refers to *"all forms of electronic-supported learning and teaching, which are procedural in character and aim to influence the construction of knowledge with reference to individual experience, practice and knowledge of the learner. Information and communication systems, whether networked or not, serve as specific media to implement the learning process".* Sun *et al.* [20] refer to e-learning as utilizing information and communication technology for delivering educational materials in real time or asynchronous manner liberating the user from time and location boundaries, while Cohen *et al.* [25] refer to it as the use of various web technologies to support education. Wiley [26] provides a broad definition of e-learning as the practice of using computers or electronic systems to support the learning process in a systematized way. According to the ASTD (American Society for Training & Development, 2001) the term e-learning covers a variety of applications and practices which include computer-based training (CBT), virtual classrooms, online learning, online teamwork and collaboration. Furthermore, e-learning refers to the use of electronically conveyed learning resources through electronic means such as the internet, satellite signals, DVDs and CD-ROMs. Effective e-learning is defined as the result of a proper combination of content (considered to be the main player in the success of building learning and knowledge) and technology, alongside services; it is used to convey knowledge and skills [27].

2.2.2 **Benefits of E-Learning**

Some of the advantages and benefits accomplished by e-learning are as follows:

- Financial savings is a critical element in the adoption of e-learning [28]. For instance, the cost of running traditional classrooms includes paying lecturers' salaries, building rentals, maintenance, student accommodation, living and travel costs.
- Reducing learning time, Brandon Hall [29] suggests that e-learning can decrease the required learning time by 40% to 60%.
- According to Fletcher [28] e-learning helps improve knowledge retention and application to a task by an average increase of 25% more than conventional approaches.
- Self-managed and unlimited time learning approach [30].
- Efficiency and convenience.
- Professional skills and knowledge attained is consistently delivered without time or place restrictions [31].
- Self-managed learning for students with different learning abilities reduces tension and increases learning goal fulfilment.
- Interactive learning engages students and improves knowledge retention.
- Availability of timely accessed and consistently updated references helps learners to master their field [32].
- e-learning when used in group scenarios has been found to increase collaboration among students [33].

2.2.3 **Limitations of E-Learning**

According to Kruse [30] there are a number of limitations and concerns for instructors and institutions:

- A huge initial investment cost is required in order to start an e-learning program

- Technology is a crucial player in the victory of any e-learning program. Technical factors such as existing technology infrastructure and its suitability to achieve instructors' as well as learners' needs, have to be considered to determine whether the required investment is justified.
- Suitability of the available e-learning materials to achieve the desired learning goals.
- Cultural and social acceptance of the use of technology as a platform for Learning
- Technophobia and unattainability of the required technologies to support e-learning [20].
- Technology portability is still unable to beat the popularity of printed workbooks or reference material.
- Lack of human interaction as well as peer-to-peer communication is a possible drawback to the adoption of technology-based learning.
- Some instructors do not possess the necessary knowledge and skills to design and build the required e-learning platforms [34].

Expanding e-learning to higher education according to Vrasidas [35] is hindered by some obstacles:

- Lack of experience in planning and designing online instructions.
- E-Learning lacks the face-to-face traditional classroom communication and that is a big concern for instructors.
- Lack of standardization.
- Lack of expertise and the required technical infrastructure.
- Instructors are overloaded by their teaching duties and do not have the time to plan and build e-learning platforms.
- There are no incentives or motivations for instructors to teach online.

2.2.4 **E-Learning Pedagogy**

Thorndike *et al*. [36] explain that pedagogy has a Greek origin and is derived from the words *"peda"* = *"child"* and "*agogos*," = *"study of"*; therefore, pedagogy is defined as the skill and discipline of teaching children. Conventionally pedagogy describes teacher oriented instruction; however, the term has recently been used to describe fine education practice (which involves *"andragogy"*). According to Nichols [37] andragogy refers to the theories and practices of adult learning. Holmes *et al*. [38] state that there is a great deal of debate about the two terms. The pedagogical teaching model is concerned with content, transfer of knowledge, educator selections and predetermining what knowledge or skill needs to be conveyed; furthermore it entails the arrangement of contents into rational units and choosing the most efficient mode for transmitting this content, for instance lectures, reading lists, lab exercises and so on. In contrast to that, the andragogical model is concerned with supplying resources as well as assisting learners to attain knowledge and skills. It also enables educators to work as facilitators, construct a set of techniques for attracting learners and detect learning needs in addition to program goals and contents. Nichols [37] insists that e-learning will only flourish by adhering to pedagogical principles and argues that it is better for the success of any e-learning program's implementation to be led by instructional designers instead of technical developers. Mehanna [39] suggests that pedagogical principles have been abandoned by e-learning specialists and as a consequence of that e-learning has been unable to reach its real potential. In a bid to overcome this issue, e-learning experts, practitioners and developers should follow pedagogical principles as they have been widely accepted and recognized in the conventional classroom environment.

2.2.4.1 ***E-Pedagogy***

Space and time are recognized by many instructors and educationalists as an asset for the conventional teaching environment. The introduction of e-learning as a complementary or additional conveyance system for old-fashioned pedagogy as an alternative to being an instrument for building novel pedagogy is still controversial. Kuriloff [40] argues that conventional learning environments provide crucial advantages for learners and instructors alike. The advantages include real time interaction, face–to-face contact, as well as peer-to-peer collaboration; however, the absence of those possibilities is considered a disadvantage of the e-learning environment. On the other hand, time and place constraints are considered a disadvantage of the traditional classroom and that is where e-learning could fill the gap by providing anywhere, anytime convenience and an effective learning solution. Kuriloff makes some recommendations for fruitful ePedagogy in e-learning solutions:

- e-learning programs should aim to build a highly positive and motivating learning environment which encourages independence and promotes high learning prospects.
- Teaching should be conducted in a collaborative manner.
- A collaborative space should be created which encourages students to work together and promotes peer reviews.
- The instructor's role in the learning process has to be redefined to meet the needs of the e-learning environment.
- E-Learning communities should be established to promote discussion between students and peer reviews to make up for the lack of face-to-face interaction of the conventional classroom.
- The freedom offered by the e-learning environment should be highlighted, in which students and educators can perform their work without the restriction of time and space whenever it is convenient to them. This is one of the most powerful and appealing features offered by e-learning.

2.3 MULTIMODAL INTERACTION

While considering the use of multimodal interaction in the computing environment, it is beneficial to reflect and ask, what does metaphor mean? According to the Oxford Advanced Learning Dictionary [41] *"the particular way in which something exists, is experienced or is done (formal)", "the kind of senses that the body uses to experience things (biology)".* According to Jaimes *et al.* [42] in computing, a multimodal system in simple terms is any system capable of communicating, using more than one single metaphor and communication channel, for instance, voice, body language, gaze, etc. Commonly used unimodal user graphical interface elements such as menus, icons and scrollbars, are not up to the potentials of modern computing capabilities and do not conform to the natural multimodal communications approach used by humans to communicate and interact. Previous research suggests that using conventional input methods mostly, using a mouse and keyboard with Graphical User interface, has dramatically reduced interaction between users and computers and confused their visual channels [43]. In addition, it has increased their cognitive load [44 and 45], that goes against the natural way that humans communicate and exchange information. Naturally humans use other communication channels and senses in conveying and receiving information [46]; they might communicate verbally using speech or non-verbally by other metaphors such as body language, facial expressions, gaze and body language. The trends of designing modern interfaces tend to be interactive and multimodal in nature. This move aims to achieve natural communication approaches between humans and computers. From the nature of the evolving interfaces, there are two main recognizable categories of interface; one focuses on alternative input means and the other is concerned with combining interface input and output methods [47].

The improvement gained from using multiple input metaphors is seen as increased usability; the weaknesses of particular metaphors are covered up by the strengths of others. For example in a scenario where a mobile gadget is used with a minor GUI and a tiny keypad, it is quite difficult to type so it would be easier to say the words instead; however in another scenario in a noisy environment it would be better to type or use a pointing device instead. Multimodal interfaces utilize human communication channels to facilitate HCI, with the aim of making them flexible, effective, efficient and transparent [47]. Multimodal systems are anticipated to be accessible, easy to navigate and win users' preference. They are capable of growing computing applications into new areas, touching people's lives and accommodating their various needs [48]. Recent research with empirical experiments has shown that the merging of visual and auditory metaphors could improve the effectiveness and efficiency of the e-learning interface [2 and 3].

2.3.1 The Cognitive Theory of Multimedia Learning (CTML)

This theory hypothesises that learning instructions built and designed in accordance to the way human cognition works will be more beneficial and help improve students' learning performance [49]. The CTML theory is constructed using three cognitive principles (for the conceptual model of CTML theory, see Figure 2-2):

a. Human procedure for processing information is a cognition process, which makes use of a dual-channelled system, the first for auditory "verbal" information, and the second for visual "pictorial, non-verbal" information.
b. Each channel is limited in its capacity.
c. Active learning takes place when the channelled information goes through a series of cognition processes, involving information coordination and integration.

The theory suggests that active learning involves a chain of cognitive processes [27]:

- The initial process starts when the student picks out images and words (auditory or visual) relevant to the presented learning topic.
- Then the relevant data is channelled from the sensory memory (through the visual channel and the auditory channel) to the operational memory where it is processed and cohesively organized into visual model and verbal model.
- The next stage is to incorporate the resulting models with preceding knowledge and experience "stored previously in long-term memory" [49].

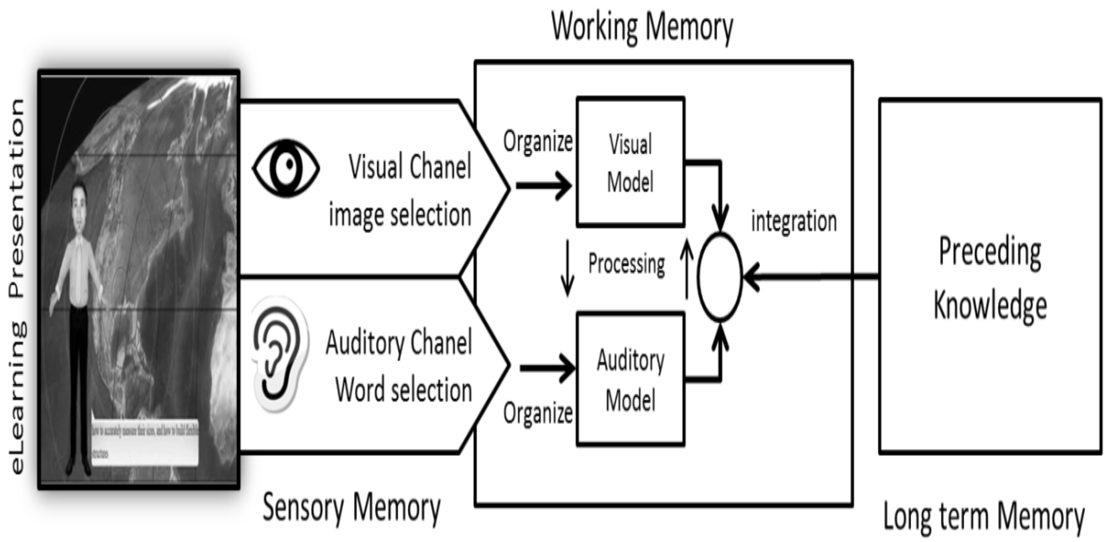


Figure 2-2: The CTML conceptual model (based on Mayer [49] work).

Adopting this theory for multimodal interaction could be useful for gaining optimal results, when presenting learning materials. Instructors as well as courseware developers should utilize the visual and auditory channels when delivering the intended knowledge and avoid flooding the visual channel as that will delay or halt the cognition process [27].

2.3.2 **Multimodal Metaphors**

2.3.2.1 ***Auditory icons***

These can be defined as concise sounds used to symbolize objects, meanings, and activities. They make use of preceding user experiences with naturally occurring sounds and their associations, they are comparable to visual icons used in graphical user interfaces to present information in an easy and simple fashion [50].

In line with the pictorial system's capacity to represent various attributes simultaneously, such as colour, form, size etc., the auditory system attributes include *"pitch*, *loudness,* *frequency, timbre and so on".* Auditory icons are abstract and easy to grasp [51]; similarly they can be mapped to encapsulate objects, events and actions to be presented directly or indirectly to the aimed event [52]. For instance, the sound of breaking glass can be used to represent an earthquake. According to Walker [53] *"The directness or auditory similarity between the icon and the actual object can vary considerably".* Provided it stimulates the sound associated with an object or act it is deemed an auditory icon. The utilization of auditory icons in computer applications is to some degree limited [54] but auditory icons are quite useful in demonstrating real environmental objects.

2.3.2.2 ***Earcons***

Earcons can be described as abstract, synthesized sounds composed by combining patterns of sounds, as well as music tones arranged in an organized and structured manner. They communicate non-verbal acoustic messages. Brewster *et al* [55] describe them as *"composed of motives, which are short, rhythmic sequences of pitches with variable intensity, timbre and register".* Blattner *et al*. [50] suggest a tree like hierarchy of earcons, with each earcon representing a node and inheriting the upper ranking earcons' attributes. Brewster *et al.* [56] propose a maximum of five ranks to this branched system; this variation is attributed to five elements:

1. Rhythm refers to the arrangements of timely sounds and silences, the division of a space of time into a defined, repeated pattern. Rhythm is the controlled movement of music in time. It may be defined as the division of music into regular metric portions; the regular pulsation of music" [57].
2. Pitch, refers to the subjective perception of the lowest and highest frequency of sound waves; it outlines the position of a tone in relation to others, resulting in the listener having the sensation of being with lower or higher frequency [57].
3. Timbre, describes the sound quality of a voice or an instrument; furthermore it can be described as the element of a tone making musical instruments sound different while playing the same musical note [57].
4. Register, refers to relative height or range of an instrumental sound or voice. Typically registers are well-defined by a variation in the sound quality among higher and lower ranges [57].
5. Dynamics is used as an indicator of relative loudness or softness (Cole and Schwartz, 2012). There are two basic signs used to indicate sound dynamicity, piano (p) = *"soft"*

to indicate softness and forte (f) = *"loud"* to indicate loudness. The signs can be doubled or tripled to indicate different degrees of softness or loudness. For instance *"pp"* = *"very soft"* and *"ppp"* = *"very, very soft"*. For moderate degrees of loudness the symbol mezzo (m) = *"moderate";* for example mp= *"mezzo-soft"* and mf = *"mezzo-forte"* [58].

Assembly, adjusting one or more of the above-mentioned earcon elements can be used to create structured and enriched audio messages. For instance, icons are utilized in some user interfaces for delivering warning messages, such as in the case of file deletion or alerting the user to a particular event. The established connection between objects and their corresponding earcons is metaphoric in nature [54]. For instance, demonstrating a process being executed, such as file deletion, is achieved by an earcon constructed using *"three-note"* notation with declining pitch and loudness. Nearly all earcons poses a symbolic relation between the information they symbolise and their sounds. Creating earcons and mapping them to particular objects still lacks standardization. Furthermore, they are merely judged by the personal preference of their composer and learning their connotations by users might require a considerable period of time.

In their study on the effectiveness of adding earcons to user interfaces, Brewster *et al*. [56] proposed a set of recommendations to improve the recognition of earcons based on their attributes:

- Recommendation for using timbre, simple tones should be avoided and different timbres should be used because they are easily distinguished [56].
- Recommendation for using pitch, pitch must not be employed as the only pointer to differentiate two earcons, as it is difficult for the audience to recognise the difference in this way. It is more practical to combine structured pitch with variety of rhythms. The maximum proposed range of pitch is (5kHz) and the minimum is in the range of (125Hz to 150Hz) [56].
- Recommendation for using register, the case of register is similar to pitch; therefore, it should never be used on its own to distinguish earcons. For optimal results, register should be combined with other dimensions of sound to attain good recognition rates.
- Recommendation for using rhythm, Patterson [59] states that sounds with comparable rhythms are most likely to be confusing to the listener. Consequently,

it is crucial to separate different earcons by making their rhythms different. Additionally, to increase recognition rate between earcons, it is better to use earcons with a diverse number of notes in every rhythm.

- Recommendation for using intensity, designers are advised to use an intensity rate of 15dB above threshold; hence, if the user alters the system's sound volume, then the sounds will not vanish [59]. This is because perceiving loudness varies from one person to another; in an auditory interface, the audience member must always exercise complete control of the sound level. Therefore, it is crucial to maintain earcons' intensity, within a close range.
- Recommendation for earcon combinations, when two earcons are displayed in sequence, it is recommended to use a gap between them; this is to help the audience tell when one earcon starts and when it ends. As recommended by Brewster, a 0.1 second interval should be used.

2.3.2.3 ***Speech***

The most natural way to present an object is through speech. On the other hand, using spoken words to introduce objects has some complications:

- Bandwidth issue, speech is known for its low bandwidth; pronounced words or expressions are more challenging to be spatially localized than a sound with a greater bandwidth such as music [53].
- Processing issue, the human brain needs to expend a considerable amount of mental effort and concentration to process speech. For instance, it tends to be quite difficult to receive spoken messages while concurrently having a dialogue with another person.
- In contrast to earcons and auditory icons, creating speech sounds is a straightforward and simple process using recorded speech or text to speech engines, such as Microsoft Speech API [60].

2.3.2.4 ***Music***

Music is a universal culture [61], the definition of the word *"music"* in English is a *"vocal or instrumental sound (or both) combined in such a way as to produce beauty of form, harmony, and expression of emotion"* [62]. This clearly links music to our emotional experiences; most human activities are associated or have emotional consequences; this is because we were designed to provoke emotions [61]. Emotions envisioned by a music composer are usually recognized appropriately by humans as young as three years of age. Music could be used by researchers to induce emotional experience in users [63]. Previous research has shown that measuring elicited positive and negative emotions triggered by music had up to 95% measurement accuracy [64]. Music can be used to induce excitement [65], change behaviour and effect the memory [66]. Listening to Mozart's orchestral music has been seen to help students with special learning needs improve their coordination skills, reduce frustration and has produced a noticeable improvement in their behaviour [67]. Kämpfe *et al.* [68] suggest that listening to background music induces a positive emotional influence and improves sport performance in adults. Adding music metaphor in the background could be beneficial to the student [69].

2.3.2.5 *User embodiment and avatars*

The Oxford Advanced Learner's Dictionary [70] describes an avatar as *"a picture of a person or an animal that represents a particular computer user, on a computer screen, especially in a computer game or chat room".* The term *"avatar"* was used for the first time in the 1980s within the virtual reality (VR) world habitat system and gained publicity through the science fiction novel *"Snow Crash"* by Neal Stephenson [71]. Since then developers and researchers of VR have produced a variety of definitions for the term. In summary those definitions briefly describe an avatar as a symbol of a user's character contained by a computerised environment [72] and is being progressively utilized in computer-based training programs [73]. An avatar acts as an interface, linking users with system resources; it could be used to present a component or a function in e-learning applications. An avatar enriches system interactivity [74]. Undoubtedly, avatars could help improve learners' interactivity and engagement [75]. Humans communicate verbally or non-verbally, using their bodies to deliver instantaneous, continuous and focused information about their behaviour, actions, presence, cognition, readiness, humour, location, personality, experiences and several other aspects in what is called body language [76]. The subject of avatars or user embodiment is evidently vital for planning and building a collaborative virtual world [77].

Types of avatar

Avatars can be classified according to their shape into humanoid and cartoon avatars [78]:

- Humanoid avatars normally resemble real human beings to a greater or lesser extent. They usually include a human face and a full human body and may or may not have similarities with the real user. When using humanoid avatars the user has a realistic and lucid exemplification in the simulated environment; they can change the way they look by altering outfits and accessories and the presentation can also have an evolving personality [79].
- Cartoon avatars are cartoon characters, which have specific cultural significance. Rather than relying on childhood cartoon figures, some adults choose cartoon avatars of a more stylish and fictional look.

Advantages of using avatars in e-learning interfaces

According to Oestreicher *et al.* [80], using avatars provides the best of face-to-face training and, combined with computer-based training, reduces the cost of human training. Avatars may offer a human touch that some students find an effective substitute for instructors who may not be available due to time or location restrictions; the following are some direct advantages of using avatars in a learning environment:

- Interaction with avatars motivates students.
- In contrast to a human tutor's limited accessibility, it is always available and not restricted by the burdens of time, space or location.
- Makes learning fun and interesting.
- Increase learners' engagement.
- Achieves higher rates of task completion and reduced dropouts.

2.4 EVALUATION METHODS

The main goal of multimodal interaction is to increase usability which refers to the ease of use of the system and should be regarded as one of the main components of the system development life cycle [81].

2.4.1 Usability

Usability as an objective of quality is well defined by the International Standard Organization (ISO) in its recommendation ISO 9241-11 as *"the extent to which a system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"* [82]. In ISO/IEC 9126 which are product-intended standards, usability is conceived as an autonomous influence to software quality and recognized as *“a set of attributes of software which bear on the effort needed for use and on the individual assessment of such use by a stated or implied set of users”* [82]. Usability insists on quality and places users and their actual needs at the centre [83]. Usability is widely considered as a vital aspect of the over-all quality of interactive e-learning applications[81]. According to Nielsen [84], It is important to realize that usability in user interfaces is commonly connected to five elements (see Figure 2-3):

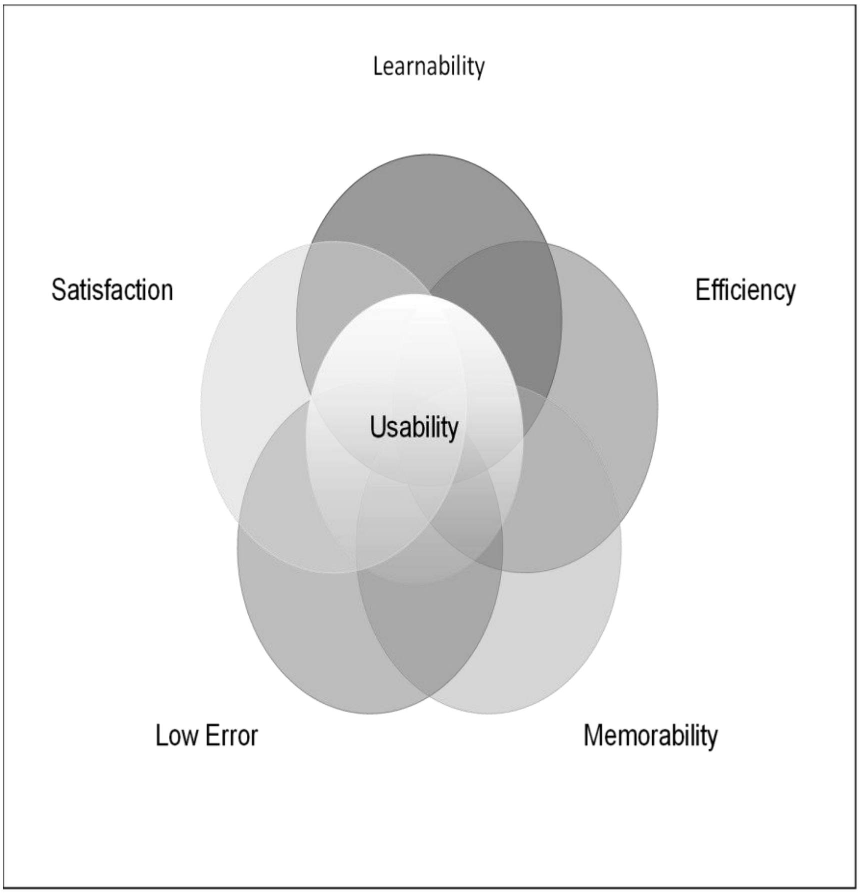


Figure 2-3: Usability Attributes (taken from Nielsen [84]).

- Learnability, for novice users of new systems, ease of learning the way that the system functions is an essential usability attribute. This is due to the fact that a user’s first experience will be learning how to use the system.
- Efficiency, once users become familiar with the system, they should be able to reach the intended productivity.
- Memorability, system functionality has to be easy to remember, so that after some time of non-use, a returning user should be able to use the system again effortlessly.

- Minimizing fault rate, the system must be capable of preventing users from making serious errors; furthermore, in case they make small mistakes, the system should be able to help users fix them without any difficulty.
- Satisfaction, to what degree users are happy with the system. They must be satisfied with it and in addition, it should increase their productivity and increase the system acceptance.

2.4.1.1 *Designing for usability*

The aim of e-learning software is to support the process of acquiring knowledge. Therefore it is crucial to consider the way scholars carry out their learning activities in addition to giving them a proper and natural means of interacting with the system as much as possible[85].

ISO 9241-11 (1998) states that *"usability is the extent to which a system can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"*[86]. Failing to recognize the importance of usability when building e-learning coursework can form obstacles as well as challenges for the learners that fall beyond the difficulty of the content or subject matters [87]. They can be occupied with learning how to use the courseware instead of learning the intended critical topics of the course [88].

Usability should be seriously considered as failing to do so leads to a poor usability design which can result in considerable negative impact on task completion rates and success scores [6]. The traditional Software Development Life Cycle (SDLC) has a number of models for this development process but usability is not considered [81] and by looking into the traditional waterfall model it is noticed that testing is left until the final stages of the development process. According to Ferré, Juristo *et al.* [89] designing for usability in initial stages leads to enhanced system usability and eliminates the necessity for conducting further tests in the final stages when it is too expensive to make drastic changes [87, 90]. Usability should be contained within SDLC (see Figure 4-2) Shows a revised waterfall life cycle. Grey boxes (1, 2 and 3) are added to include usability, while white ones refer to the classical model. Adopting the usability testing techniques in SDLC within the early stages in software developments increases efficiency, satisfaction and subsequently productivity [89].

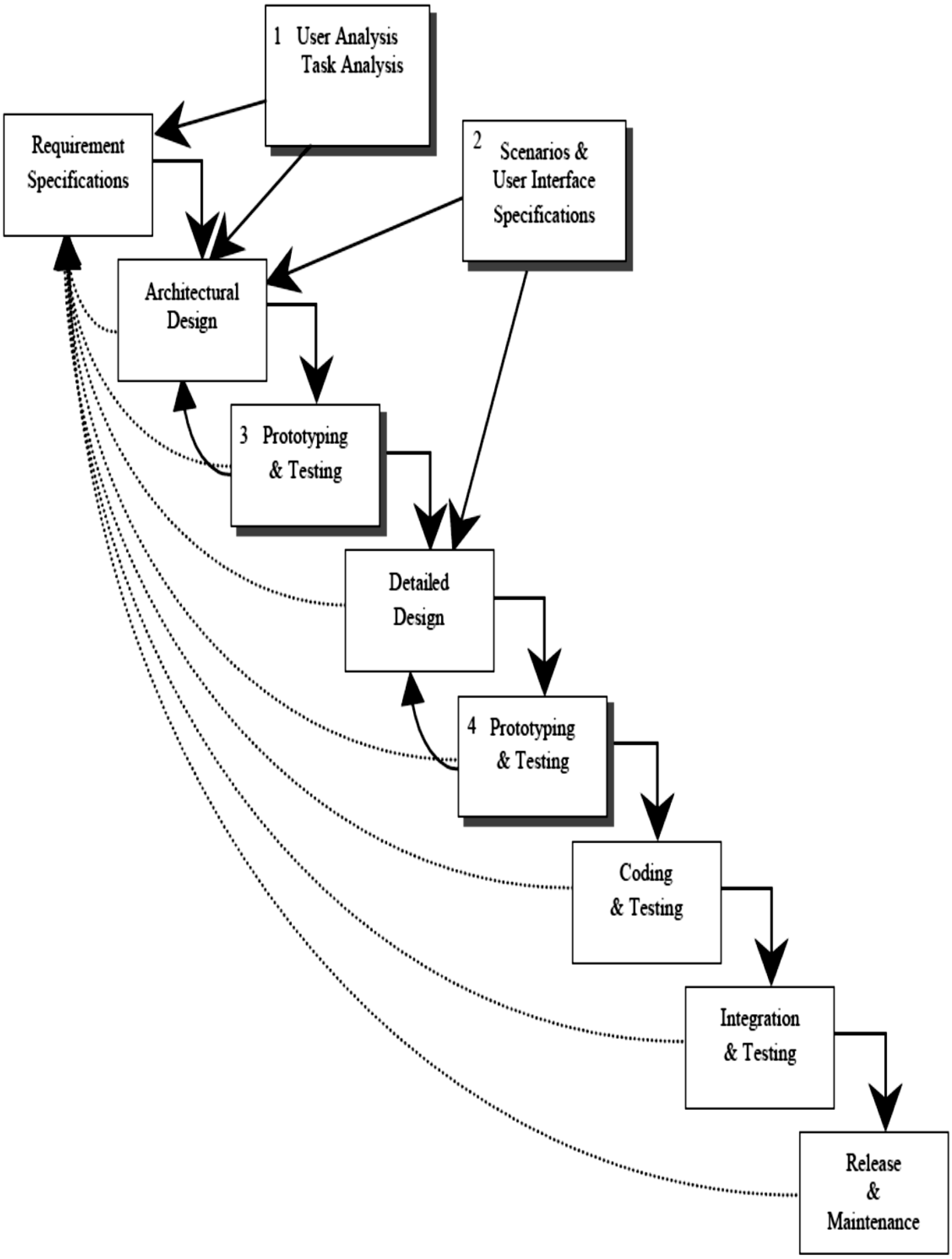


Figure 2-4: Usability and the waterfall model (taken from [81]).

2.4.1.2 *Usability Evaluation*

Usability has a crucial part to play in carrying out the successful implementation of e-learning applications[91]; therefore it is vital to conduct a usability evaluation test.

Testing aids in finding the usability level of the interactive application and determining whether the design works and lives up to its promises [89]. The context of system real use should be identified in order to be able to carry out valid and accurate testing [82]. The usability evaluation process normally makes use of qualitative data from user comments combined with quantitative data and compares them with the formerly pre-set usability objectives [92].

2.4.2 **User experience**

User experience (UX) is a new term being introduced for system evaluation [8]. It is seen as a result of computing involvement in all aspects of everyday lives; users expect more than ease of use or achieving a particular task, they expect enjoyment and entertainment. The term UX, was initially introduced by Norman [9, 93], it has evolved to extend far beyond the conventional usability measurements of effectiveness, efficiency and the traditional concept of satisfaction. The definition provided by ISO for user experience is *"a person's perception and responses that result from the use or anticipated use of a product, system or service"* [10]. UX is gaining increased interest even though there is some confusion between the UX and usability [94].

2.4.3 **Usability vs User Experience**

While some experts have tried to differentiate between UX and usability, others have tried to eliminate the borders between them. In both theory and practice user experience and usability are interrelated and usability is incorporated in user experience evaluation methods [95]. While usability is task performance oriented, UX is concerned with lived experiences [96]. Usability evaluation methods are objective and focus on measured results such as task completion time and the number of correctly-performed tasks and those which do not satisfy the subjective UX [97].Despite the fact that satisfaction in usability is subjective in nature and perceived as part of UX, it is only a small pixel in the big picture of UX, which responds to a wide range of other subjective merits such as attractiveness and hedonic and pragmatic qualities. For instance the user's enthusiasm and anticipation makes a greater contribution to UX than in the conventional concept of usability [98]. Dillon [6] argues that it is clear that usability on its own, at least as it is currently understood, is inadequate for ensuring high-quality user experiences with a new technology.

Beyond Traditional Usability Concepts		
Aspects	Usability	UX
Holistic	Usability in outlining perspective of use stresses performance in addition to user satisfaction and their accomplishment.	UX adapts a holistic approach, aiming at balancing two aspects of the system: • Task-focused approach (concerned with pragmatic side of the system). • Non-task focused approach (concerned with the hedonic side of the system) including aesthetic aspects, stimulation, attraction, beauty, challenge, and expressing user identity.
Subjective	Usability driven by objective measurements of tasks: • Effectiveness (proportion of correctly performed). • Efficiency (times and error).	UX is more motivated by users' subjective feelings and their perception of interacting with the system.
Objective	Usability uses a problem solving approach for improving them.	UX methodology of improvements lies in paying more attention to the positive context of use and how to capitalize on them to make the use of the system a happy, enjoyable, and engaging experience.

Table 2-1: User experience beyond traditional usability concepts [99].

For delivering the desirable UX, Dillon [6] Insists on addressing three key elements for successful user interaction with the system:

a. Process, what are the user's actions and practices to accomplish a particular task? For instance, the frequency in which a particular feature such as help and system menus were utilised. As a result, the user's habits as well as the system's responsiveness and technical hitches will be understood.
b. Outcomes, what are the user's intended accomplishments. For instance, what defines the objectives achievements at the end of user interaction with the system? The answer, from a user angle, constructs a solid understanding of the meaning of fulfilled accomplishment.
c. Affect, how does the user feel about the system? This embraces the perception of satisfaction from the description of usability, but then again exceeds that to take account of all users' feelings, emotional responses and interactions. Users might feel in control, irritated, entertained or self-assured. This develops recognition of users' affective interface with the system.

Bevan 2008 [11 and 12] proposed extending the definition of user satisfaction to include these elements:

a. Likability: to what level the user is pleased with their attainment of practical aims and significances of use.
b. Pleasure: to what extent the user is fulfilled by the attainment of hedonic objectives of stimulation [13] and accompanying emotive reactions.
c. Comfort: to what extent the user is comfortable with the system.
d. Trust: to what extent the user is confident that the system will perform as anticipated.

In Figure 2-5, the columns are the quality characteristics that contribute to the overall user experience needed to achieve these qualities.

The ISO describes UX as: *"a person's perception and responses that result from the use or anticipated use of a product, system or service"* [10]. UX differs greatly from usability because of its emphasis on the following factors [14]:

- Positive aspects of the user system relationship (for instance, enjoyment rather than frustration).
- The incorporation of hedonic (non-instrumental) aspects.

- UX emphasises the subjective side of system use (rather than objective performance criteria, which is the focus for usability).

Although UX differs substantially from usability, Bevan proposes the integration of UX and usability in user-centred design development. More attention needs to be paid to user's selections and preference, aesthetic feeling, frustration and sense of accomplishment [6].

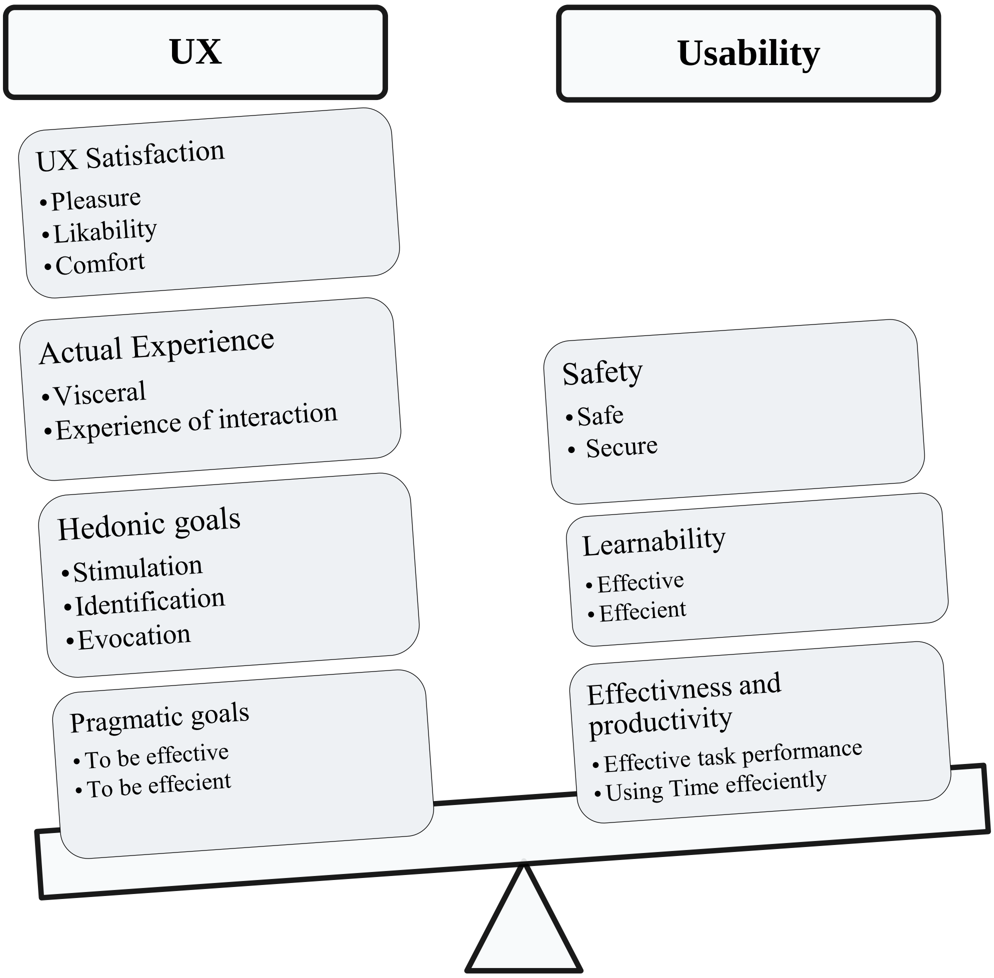


Figure 2-5: Usability and user experience attributes (adapted from [14]).

2.4.4 User Affective State

Human are emotional being and they are influenced by their contact with the surrounding environment, people, situations, learning skills, performing tasks, interacting with computer program or using a website. In addition, it is reflected upon their thinking attention, perception and their memory [100]. Users affective (emotional) state is made up of two factors, psychology and physiology; the most common conceptual model for understanding the users affective state is PAD (Pleasure-Arousal-Dominance) model based on the work of Mehrabian *et al.* [101] and Bradley *et al.* [16]. PAD consists of three fundamental dimensions, pleasure (pleasant valence vs unpleasant valence), arousal (relaxed vs aroused) and dominance (demotivated vs motivated) [102] and is widely used concept within the field of computing [103].

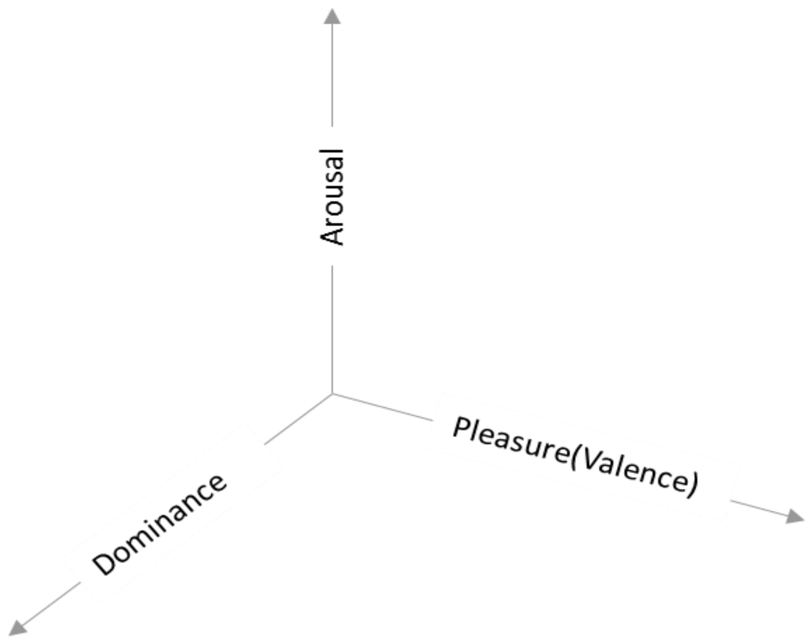


Figure 2-6: PAD model (based on Mehrabian *et al. [101] work).*

Valence: is the evaluation of a subjective experience of emotion ranging from *"positive"* to *"negative"* state [104] , provoked by the surrounding environment, or subjective feelings [105].

Arousal: The term arousal in physiology and psychology refers to the state of being active when presented with a stimulus ,arousal indicates the strength of particular emotion [106] and results in an increased skin conductivity[107]. Arousal might be positively or negatively valenced according to the way in which they are perceived, for instance joy is associated with high arousal and valence while anger is associated with higher arousal and low valence. Arousal can be measured either subjectively or objectively, as it results in psychophysiological changes such as increased skin conductivity [108].

Dominance: The term dominance can be viewed as the extent to which users assume control over their reaction to move backward (are demotivated) or forward (are motivated) when presented with stimuli ,it can be described as a dominance-submission of a person in a given situation [109]. Motivational intensity refers the strength of urge to move toward or away from a stimulus.

2.4.4.1 *What is an emotion?*

According to the free encyclopaedia the term emotion in psychology is *"a subjective, conscious experience characterized primarily by psychophysiological expressions, biological reactions, and mental states"* [110].

The origin of the word emotion goes back to the 16th century and is derived from the French word *"émotion",* the root of the word is Latin *" emovere, from e- (variant of ex-) 'out' + movere 'move' "* [111]. The word *"emotion"* and its connotation have changed over the years in line with the changes in theories of emotion [112 and 113]; there have been many theories and models which have attempted to explain emotions and how they originate [61].

The most prominent theories, which have attracted many scholars' attentions, are:

The James–Lange Theory: W. James and C. Lang stated that emotional stimuli trigger bodily (somatic) arousal, which happens unconsciously [114]. Then the brain in response, interprets those changes as a state of emotional feeling and physiological changes[115] [106] (see Figure 2-7A ,for a conceptual model of the theory).

The Cannon–Bard Theory: W. Cannon argued that, somatic changes due to a reaction to external stimuli are not precise enough to produce physiological changes and particular emotions. Following this, P. Bard proposed that, emotional and physiological responses are parallel and occur at the same time. This resulted in what is known as the Cannon-Bard theory, which proposes that, in observation of emotive stimuli, the brain answers instantaneously. However, separately, bodily changes and subjective feelings of emotion are prompted [106] (see Figure 2-7 B, for a conceptual model of the theory).

The Singer-Schachter Theory: This theory, also known as the two-factor theory, proposes that emotions are the products of a combination of physiological arousal and cognition process. The theory indicates that when a bodily arousal takes place, a cognition process occurs to interpret the meaning of the arousal in relation to the situation in which that particular arousal has occurred. Cognition is affected and influenced by the surrounding environment and is reflected in the personal emotional state [116 and 117] (see Figure 2-7 C for a conceptual model of the theory).

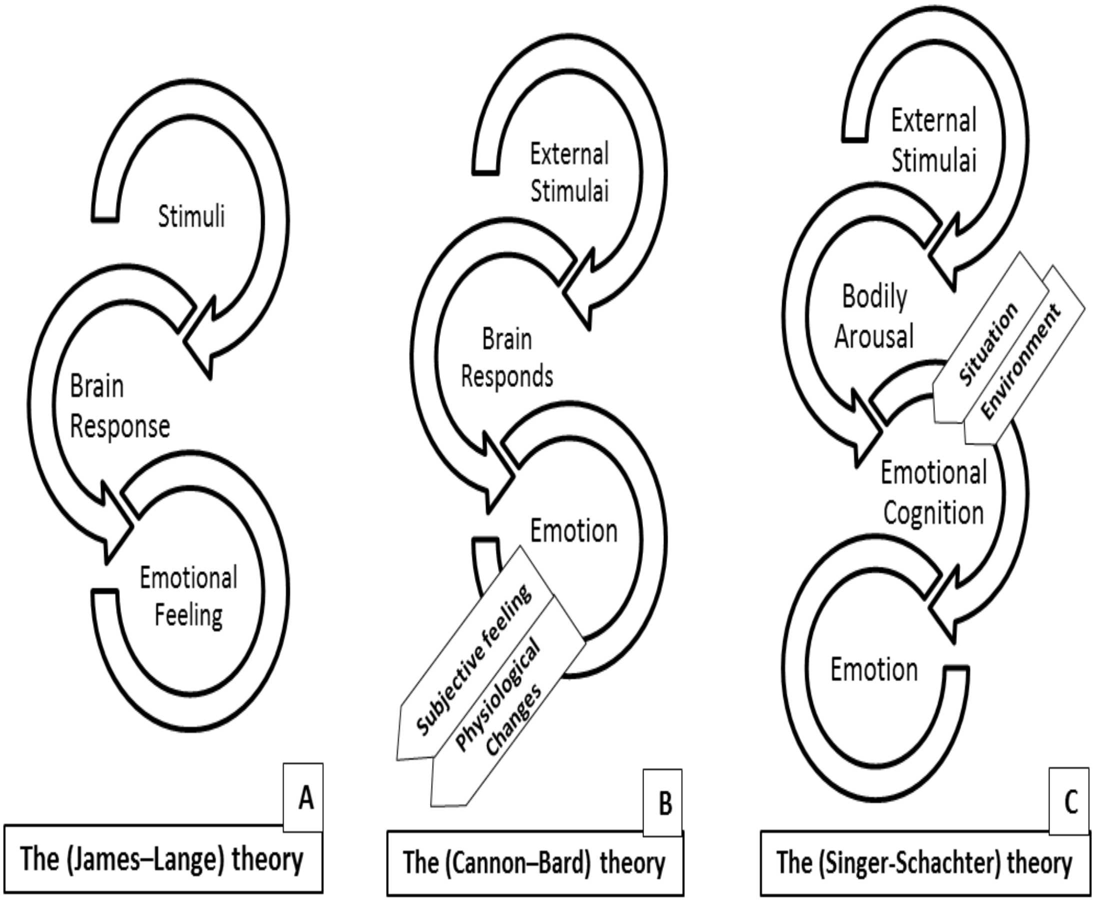


Figure 2-7: Theories of emotions.

2.4.4.2 *Affective States (Categories and Models)*

There has been a great deal of debate on how to categorize or classify emotions due to the variability and nature of the subject. Scientists have proposed many models to classify emotions. The most commonly used models are the Basic Emotion model [118] and the Two Dimensional model [119]. The basic emotion model was based on results of Ekman's experimental work. Ekman proposed that, there are discrete and universal basic emotions shared across all cultures. The primary emotions are *"surprise, happiness, anger, fear, disgust and sadness"* [120] (see Figure 2-8). Furthermore, Ekman states that the more complex emotions are secondary to these basic ones and are built on them [118, 121]; The basic emotions were selected because of their distinctiveness in their physiology, appearance, duration, occurrence, memories and subjective experience [122]. The two-dimension model by Bradley and Lang [119, 123] places emotions into two dimensions, space of valence and arousal. The model categorises emotion according to their valence *"pleasantness"* and arousal *"intensity"* (see Figure 2-9 A).

Even though several attempts have been made to categorise emotions differently, the Basic Emotion model has a great deal of agreement among scholars in the field of psychology [124] (See Figure 9-2 B, for the demonstrations of the elementary emotions portrayed on valence-arousal space) [63]. For instance, sadness has a negative valence and low arousal (see Figure 2-10).

Figure 2-8: Pictures used in cross culture studies (taken from Ekman[125]).

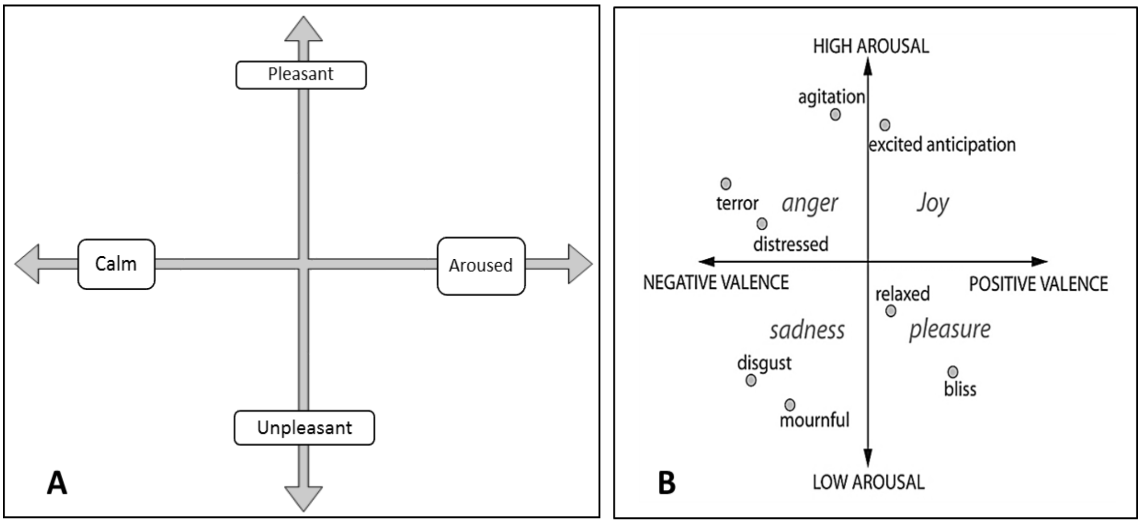


Figure 2-9: (A) Two-dimension space of valence and arousal

(B) Primary Emotions (taken from Kim and André [63], Bradley and Lang [126]).

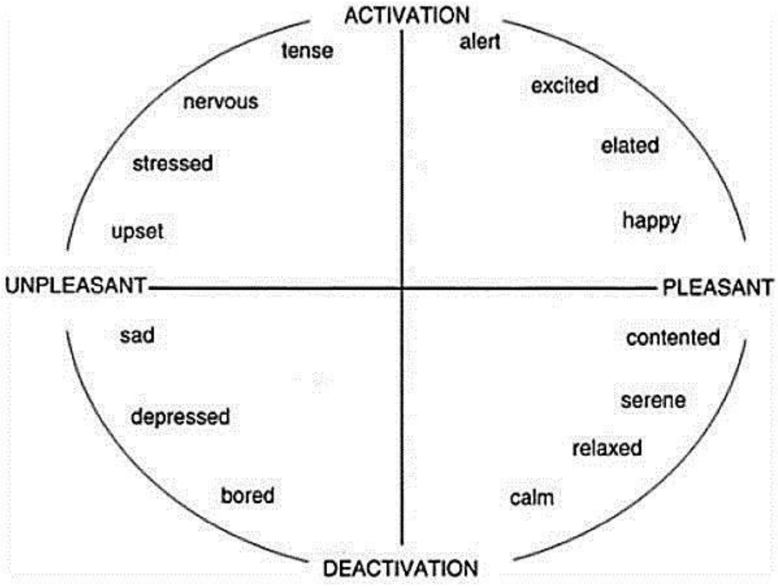


Figure 2-10: The Circumplex model (taken from Posner [127]).

2.4.4.3 *The Relationship between Affective State and the Nervous System*

The nervous system is divided into two key branches [61] (see Figure 2-11 for an overall view of the nervous system):

- The central nervous system (CNS), branches into the brain and the vertebral cord.
- The peripheral nervous system (PNS) links the CNS to the body's organs and limbs.

Furthermore , the CNS sends and receives signals from the PNS [128], in sequence the PNS splits into two subdivisions [129]:

- The somatic nervous system (SoNS) is responsible for controlling the body's intentional movements) [61].

The autonomic nervous system (ANS) which is known as well as the "involuntary" nervous system is linked to human emotions [130] and is also divided into two branches:

  - o The first branch is the parasympathetic nervous system (PSNS), which is also known as the rest-digest system and is accountable for controlling the body during resting.

- The second branch is the sympathetic nervous system (SNS) which is also recognized as the flight-fight system and is activated due to emotional arousal [131]. It has been a long time since scientists first utilized physiological changes to measure the intensity of experienced emotions; the first attempt to record human autonomic nervous system activities [132] dates back to the late decades of the nineteenth century.

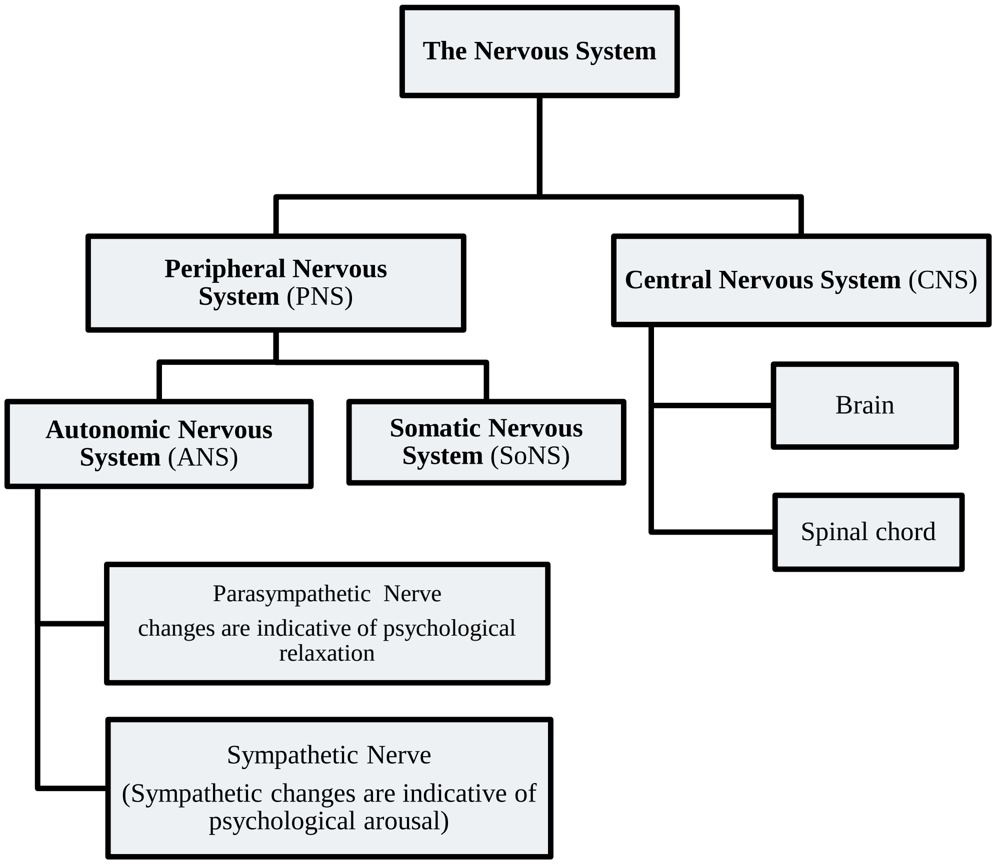


Figure 2-11: The nervous system, an illustration diagram.

Physiological signals being used to note consumers' preferences and behaviour is an emerging area of research known as "Neuromarketing" [133]. When using physiological signals to study effects of external stimuli such as arousal, one should consider using skin conductance (SC) measurements in particular. That is due to the fact that, SC change is due only to the activation of the sympathetic branch of the Autonomic Nervous System (ANS); this is because changes in other physiological signals such as blood pressure, heart rate and pupil diameter could be caused by the activation of either parasympathetic or sympathetic ANS branches. Another advantage is that SC changes are noticeable even from a single stimulus in comparison with other signals [134]. Research methods which adopt self-reported emotional experience techniques are merely dependant on the accuracy of the information provided by the participants and how truthful they are in their responses and feedback [135]. Furthermore, sometimes users experience emotions but they do not report them simply because they are unaware of them; this is known as unconscious emotional experience. Berridge *et al.* [136] and Winkielman *et al.* [137] argue that there are indications that negative as well as positive affect can be subliminally provoked and stay hidden from any subjective feeling and regardless of that obscurity subliminal emotions have great impact on our likes and the choices we make.

2.4.5 **User Experience and Affective State**

Acquiring knowledge is an experience; it is merely a cognitive process, in which learners are affected by their emotional state. For instance, successfully completing a test could induce a positive attitude and a learning experience; in contrast, failure could have the opposite effect [138 and 139]. In psychology, affect can be described as an emotion or feeling of subjective experience [140]. Emotions play an important role in shaping UX and making it a positive or negative one. Furthermore, users' emotions or affective state when using any system will determine how they perceive the hedonic as well as the pragmatic qualities of the system's performance and preference. How are user experience and affective state measured? Scholars studying user experience and emotions typically use many techniques, including self-reporting, observation and biofeedback. The techniques used can be classified into two categories, subjective and objective, according to their nature.

2.4.6 **Subjective Measurements and Evaluation Methods**

These methods are purely dependent on users' reflections on their experience with a system or a product. The following are some of them together with their advantages and disadvantages.

2.4.6.1 ***Expressing Experiences and Emotions (3E)***

This emotion-measuring technique, is known as the 3E method and was proposed to collect users' emotions and experiences [141]. The 3E technique is suitable for evaluating an application or a product. Its layout consists of a stick figure with a clear face in addition to speech and thought bubbles; the face does not show any emotions (see Figure 2-12) users may use them to express their feelings and experiences in words or non-verbally in drawing. Although users can express their feelings without restrictions with 3E, analysing the user’s output is time-consuming and errors in the interpretation of drawing and expression are more likely to happen.

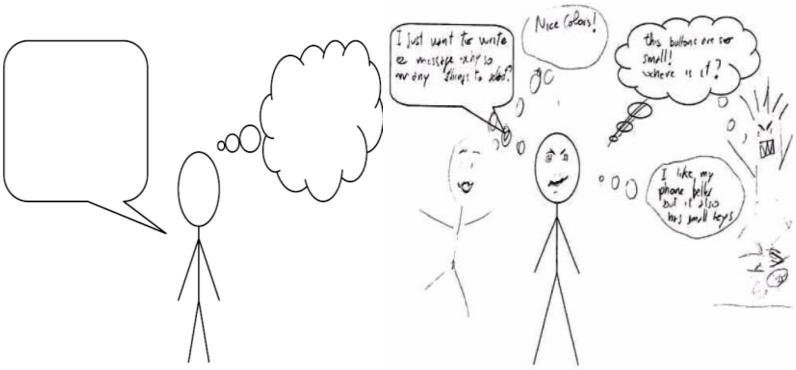


Figure 2-12: The *“3E”* layout (taken from [141]).

2.4.6.2 ***Affect Grid Scale***

Affect Grid is a two-dimensional scale with 9x9 grid (arousal is represented by the y-axis and pleasure by the x-axis). It is intended as a quick way of measuring affect at any particular point in time [142]; with the affect grid, users can mark their emotion at any time during their course of interaction with an application or a product (Figure 2.13). The scale is simple and easy to use, however it does not provide many details and it is not extensively validated.

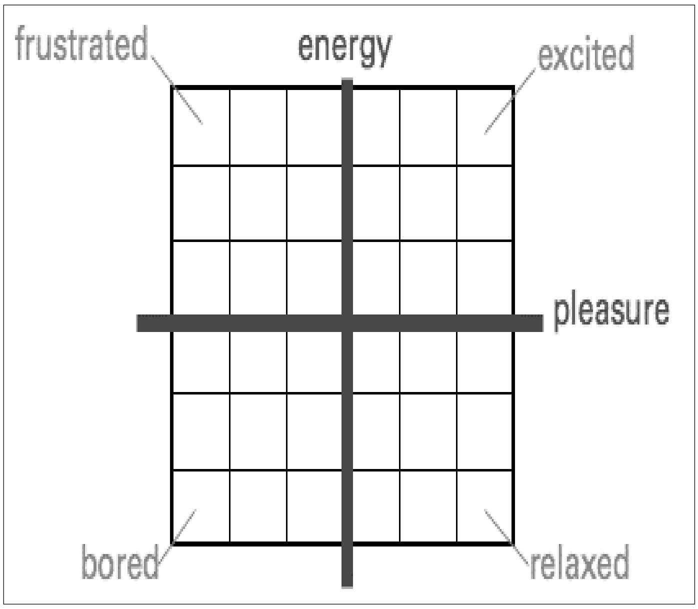


Figure 2-13: Affect grid (taken from [142]).

2.4.6.3 *Emocards*

Emocards are a self-report instrument which assists users in expressing their feelings about a task or a product [143]. This non-verbal technique is presented as set of cards or printed on a white paper (Figure 2-14). Upon the completion of the task, the user picks a single cartoon face to express their own emotion. Emocards has the advantage of being cheap to use and does not need special skills to administer or interpret, however, it is limited to only eight categories. It also does not provide further details about the emotion felt or the affective state.

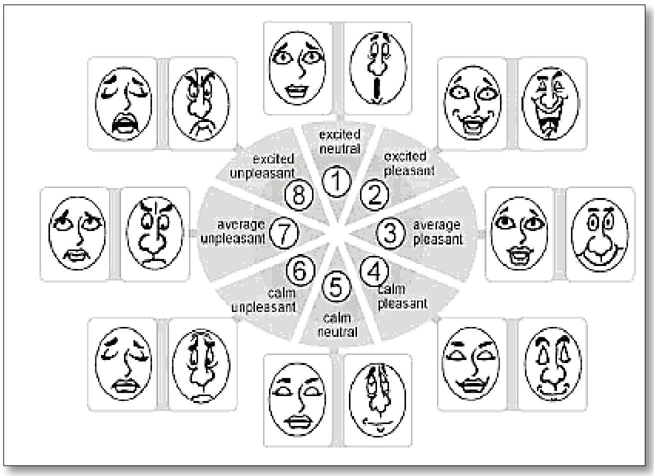


Figure 2-14: Emocards (taken from [143]).

2.4.6.4 *Sensual Evaluation Instrument*

While interacting with the system, sensually shaped objects (see Figure 2-15) are utilized by the user to express how they feel. When an emotion is stimulated, they can pick any shape to present it. The test is followed by an interview to interpret the test outcomes [144]. It is not distracting to use while interacting with the system like other self-report forms; however, it is not an easy task to analyse the data and interpret results.

Figure 2-15: Sensual evaluation instrument (taken from [145]).

2.4.6.5 *Positive and Negative Affect Scale (PANAS)*

PANAS is a self-reporting subjective scale which was created to measure two extents of dimensions (positive and negative affect) [146]; it was originally developed for clinical trials, however it could be used to evaluate affected users' moods in other evaluation settings [147]. Unfortunately, PANAS is a lengthy verbal scale with a lot of wording, which makes it difficult for users to understand, and this might hinder their experience and have a negative effect on their emotions and rating of the tested condition.

2.4.6.6 *Geneva Emotion Wheel (GEW)*

Geneva Emotion Research Group Developed the wheeling system to acquire self-reported insight into sensed emotions provoked by an event or a product [148 and 149]. Using a wheel-shaped scale, the user can identify the emotion felt and choose its intensity (see Figure 2-16). A well-established research group designed the tool; however, the subjective scale does not provide any further details about the emotion felt and the holistic user experience.

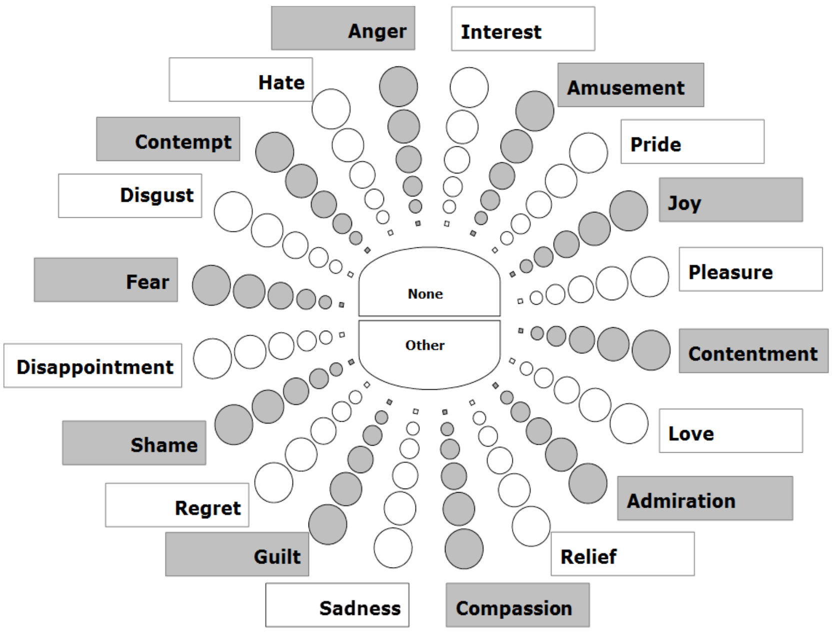


Figure 2-16: Geneva emotion wheel (taken from [148]).

2.4.6.7 *iScale*

iScale is a tool for recalling a user's emotions and experience with a product or an application over a period of time [150]. Users draw curves to describe their experiences and feelings about the tested item over a period of time. The tool can assist in revealing users' experiences over time without interfering with the user during the course of interaction; however, there is a major issue of concern, iScale solely depends on the user's ability to remember events rather than using a real-time approach (See Figure 2.17).

Figure 2-17: iScale (taken from [150]).

2.4.6.8 *PrEmo*

PrEmo is a non-verbal self-report instrument that is used to recognise 14 different types of emotion, which are depicted by animated expressive cartoons (See Figure 2.18). With PrEmo instrument, users non-verbally express their feelings toward the presented product [151 and 152]; the instrument is suitable for cross-cultural studies and users can select more than one character to describe mixed emotions. However, PrEmo does not provide a deep insight into the emotions felt and the whole user experience, concentrating only on the aesthetic aspect of the tested item.

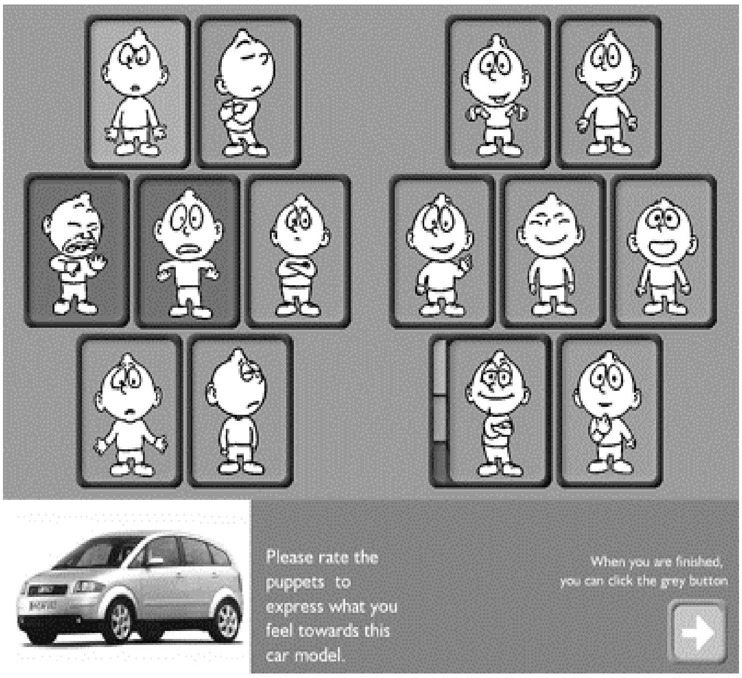


Figure 2-18: PrEmo (taken from [153]).

2.4.6.9 *Experience Interview*

Interviews are normally used in empirical research [154]. They are a good way to collect qualitative data when it is not possible to use quantitative or when it is not sufficient. They can be conducted face to face or online where users are asked how they feel about a situation or a product. According to their format, interviews can be divided into three types [155]:

1. Structured Interview, this type of interview adheres strictly to a fixed set of scripts and questions. The interviewer must be consistent at all times in the way they conduct the interview and avoid reacting to the interviewee's answers; this type of interview is deemed useful when the interviewer is seeking very specific information.
2. Semi-structured Interview**,** these are less formal than structured interviews; the interviewer must adhere to pre-set questions, however they are allowed to explore participants' answers and seek further clarification. This type of interview provides a deeper understanding of the interviewee’s answers.
3. Unstructured Interview**,** the style of the interview is more like a friendly chat. This type of interview is used when the researcher is trying to gather as much information as possible about the area being studied. Although interviews are considered to be a source of data, they are demanding, time consuming and require the researcher to be experienced in conducting such interviews; otherwise they are very costly and might be misleading [154].

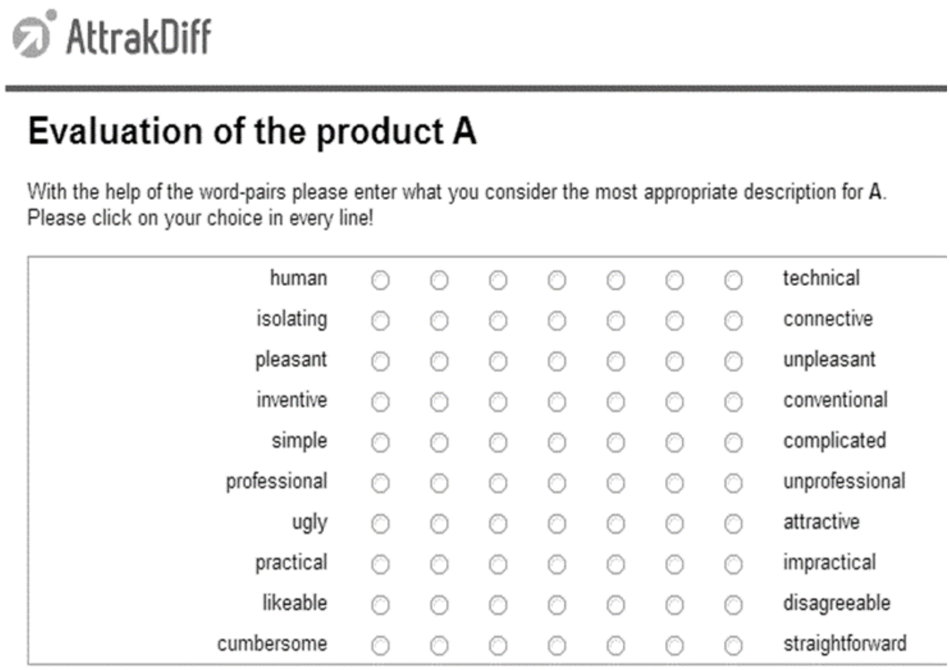
2.4.6.10 *Emotion Sampling Device (ESD)*

Emotion Sampling Device is a PDA running a java application. The application provides a sequence of queries to collect the user’s emotions experienced as consequence of a particular event [156]. ESD-answered questions may recognize up to seventeen different emotions. The ESD cannot be used in the early development stages and it is not available in the public domain.

2.4.6.11 *AttrakDiff*

AttrakDiff can be described as a tool for measuring user experience with an interactive system or a product [15]. Using pairs of opposite adjectives, users may indicate their experience (using a 7.0 point scale ranging from -3 to +3); from users’ selections it is possible to evaluate the desirability of a product in terms of functionality, appearance, attractiveness and whether further optimization is needed (see Figure 2-19). AttrakDiff utilizes three dimensional assessment strategy [157]:

1. First dimension, Pragmatic quality (PQ) investigates the aspects of functionality and detects whether the user is able to achieve the anticipated goals from using the product.
2. Second dimension, Hedonic quality (HQ) splits into two sub-divisions:
    a. Hedonic Quality - Identity (HQ-I) finds to what degree, users have identified or been led into a particular identity by the product (for instance, leaving an impression of being professional, calm, contemporary).
    b. Hedonic Quality - Stimulation (HQ-S), individuals have urgency to advance and progress. Hence, the stimulation side of HQ is used to evaluate to what extent the user is stimulated by the product. The stimulation occurs due to uniqueness, motivation, inspiring utilities, content, performance, and appearance delivered by the product.
3. Third dimension**,** Attractiveness (ATT) refers to subjective assessment of a product's attraction. Attractiveness rating reflects an overall ranking based on perceived quality [14].


AttrakDiff

Evaluation of the product A

With the help of the word-pairs please enter what you consider the most appropriate description for **A**.
Please click on your choice in every line!

human	○	○	○	○	○	○	○	technical
isolating	○	○	○	○	○	○	○	connective
pleasant	○	○	○	○	○	○	○	unpleasant
inventive	○	○	○	○	○	○	○	conventional
simple	○	○	○	○	○	○	○	complicated
professional	○	○	○	○	○	○	○	unprofessional
ugly	○	○	○	○	○	○	○	attractive
practical	○	○	○	○	○	○	○	impractical
likeable	○	○	○	○	○	○	○	disagreeable
cumbersome	○	○	○	○	○	○	○	straightforward

Figure 2-19: AttrakDiff (taken from [15]).

2.4.6.12 *Self-Assessment Manikin ((SAM) Scale*

SAM is a self-report scale for the assessment of emotions. It was constructed based on PANAS [147] and measures three dimensions of emotion (valence, arousal and dominance) [16]. SAM is a 9-point scale, which utilizes cartoon characters to exhibit emotions (see Figure 2-20) and is used after users have completed their trial with the tested condition. It is common for SAM to be used in valuing advertisements and increasingly in product evaluation. As SAM is a non-verbal tool, it can be used in a cross-cultural situation for evaluation purposes. SAM users are able to rate their emotions, by selecting one of the nine characters in Figure 2-20 (A), rate its intensity by selecting a character from Figure 2-20 (B) and their control of the situation by picking a character from Figure 2-20 (C).

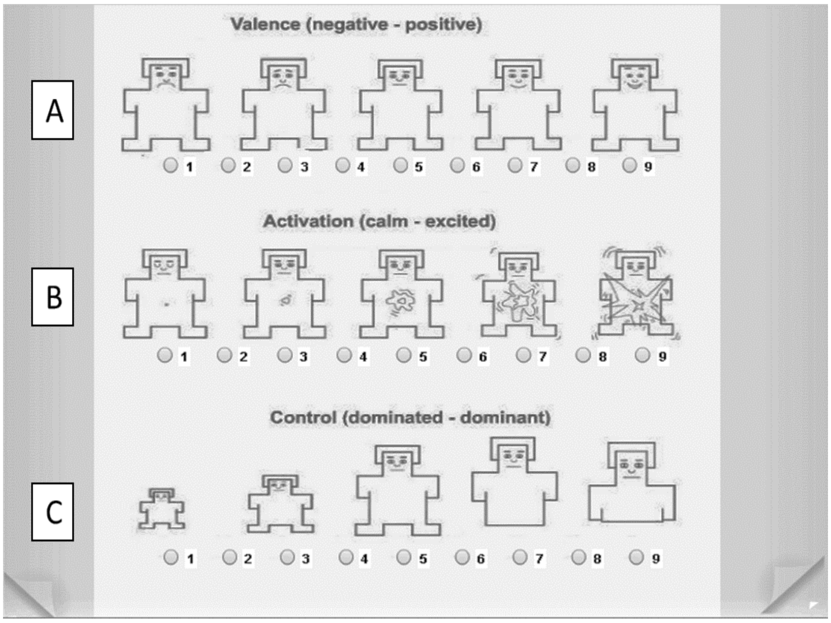


Figure 2-20: SAM self-assessment scale.

2.4.7 Objective Measurements and Evaluation Methods

Research methods adopting self-reported emotional experience techniques are solely dependent on the accuracy of the information provided by the participants and how truthful they are in their responses and feedback [135]. Furthermore, sometimes users experience an emotion but they do not report it simply because they are unaware of it; this is known as unconscious emotional experience, Berridge [136] and Winkielman [137] argued that there are indications that negative as well as positive affect can be subliminally provoked and stay hidden from any subjective feeling. Regardless of that obscurity, subliminal emotions have a great impact on our likes and the choices we make.

2.4.7.1 *FaceReader*

This tool is a real-time tracker of the user’s affective state while interacting with an application or a product [158 and 159] (see Figure 2-21). Using a video, face reader is able to identify six emotions (joy, anger, surprise, sadness, fear and disgust) from facial expressions at any time.

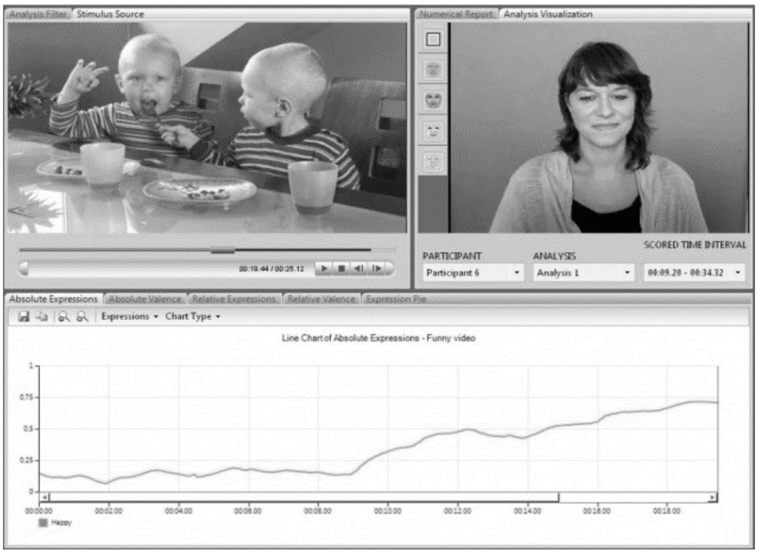


Figure 2-21: FaceReader, emotion recognition (taken from Noldus [159]).

Software vendor, claims up to 90% success rate in emotion recognition from facial images. Automated emotion recognition is quite helpful in saving time by speeding up the process of analysing facial expressions and avoiding mistakes made by novice or non-expert researchers. Another advantage is that there is no interruption to the user while using FaceReader; however, its ability is limited to recognizing only six emotions. Furthermore, it cannot be used in the early development stages, only with a finished product.

2.4.7.2 *Psychophysiological Measurements:*

Psychophysiology is an area of psychology which specializes in utilizing the body’s physiological processes to study its psychology [64]. The main advantage of physiological measurements is that even if the user does not express emotion through subjective or bodily gesture, change in their physiological signs is noticed; this is due to the involuntary stimulation of the SNS [160] as a result of the person experiencing positive or negative excitement [63]. The activation of the ANS is seen as change in heart rate, blood pressure, respiration speed [161], changes in electro-dermal activity (EDA)

[162] as well as pupil size and pulse; these changes could reveal the user's emotional state when interacting with a system or product [163]. Furthermore, psychophysiological changes cannot be caused intentionally nor consciously controlled by humans. Therefore hiding physiological signals triggered by emotions or faking them is not possible [64] Additionally, physiological signals allow us to recognize subtle positive and negative affect variations which are not apparent or visually observed [164 and 63].

2.4.7.3 *Commonly used psychophysiology measurements:*

The type of psychophysiological measurements depends on the source of measured signal (See Figure 2-22).

Cardiovascular system measurements: Cardiovascular measures use records of heart activities to identify positive and negative emotions. Heart Rate Variability (HRV) is the most common measurement within this group; HRV is the alternation rest between two successive heartbeats. It is helpful in evaluating and showing stress levels in adults [63]. However, one issue to consider when using HRV is that, two main branches of the autonomic nervous system (SNS and PSS) affect the heart. There is no way to identify whether the source of any collected data is due to an emotional arousal or to another physiological process. Therefore using cardiac measurements might result in misleading conclusions.

Respiratory system measurements: Respiratory System Measurements portray how deeply and steadily an individual is breathing and are interpreted as an indicator of negative valence. However, they are not appropriate for virtual applications because of the slow signal of the physiological reaction [165].

Brain measurements: Known as Electroencephalography (EEG), these signals point to the activity of the CNS [166] [167] yet they are not well-matched for real-world applications due to their high sensitivity to bodily movements; for instance, blinking of the eyes or an electrostatic propagation because of the use of a large quantity of probes [168].

Electro-dermal activity (EDA): The skin is the largest body organ and is innervated by the SNS [126]. This makes skin activity-captured data the best candidate and the most cost effective way to measure the "arousal" dimension of emotion. EDA refers to the recorded variations in the skin's capability of conducting electricity. The measurements reflect the overall arousal of the SNS due to external stimuli [169]. Changes in the EDA occur when the brain sends a signal to the skin's sweat glands to increase sweating; humans might not notice or feel any sweating on the skin, however the changes in skin conductivity are significant and quantifiable [17]. The EDA signal can define the amount of an individual's excitement or nervousness, as well as indicating increased somatic activities or arousal fluctuations linked to emotional feelings such as pain, happiness, disgust, reasoning and many others [170].

There are a number of ways to measure EDA, including skin potential, skin resistance and skin conductance [162]; however, following the guidelines and recommendations made by the Society of Psychophysiological Research [171], the skin conductance (SC) technique has become the universal standard practice to study the EDA [172]. The preferred method for the technique is to measure skin conductivity by passing a small current through the skin between two electrodes [173]. Changes in skin conductance indicate ANS activation and emotional arousal.

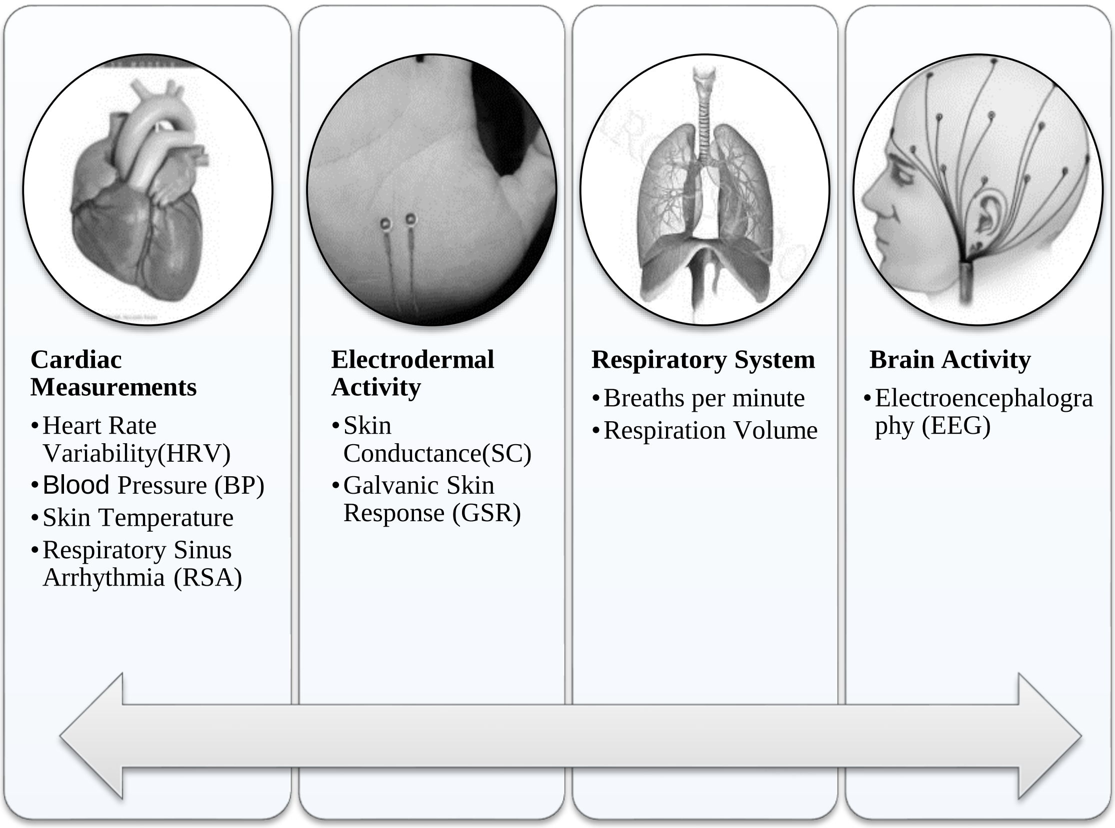


Figure 2-22: Types of physiological measurement.

2.5 RESEARCH TRENDS

There has been no previous research, which empirically evaluated user experience and affective state in relation to the usability of multimodal e-learning interfaces and in particular to investigate multimodal communication metaphors and compare them. However, there have been studies in the discipline of human-computer interaction and in the medical field using physiological signals to recognize emotions, including pleasure, sadness, engagement, excitement, frustration, anger, boredom, ease and many more. Cannon [174] studied the relation between what he described as bodily changes and their relation to emotions such as fear, pain, hunger and rage. Newton, *et al*. [175] investigated the relation between peripheral blood circulation and emotion. Changes in the affective state are accompanied by changes in ANS, which is reflected as change in blood circulation in skin vessels and results in alteration of the skin temperature. Two groups of 14 nursing students participated in a trial and the group that was subjected to the negative stimuli had lower mean values of finger skin temperature in comparison with the group, which was positively stimulated. This shows that changes in skin temperature are linked to the affective state. There was a study by Carpenter and Haddan [176] investigating the relation between students' physiological activities and their scores to establish connections between students' learning achievements and increased physiological skin signal. Students with greater physiological skin signal for the duration of learning scored best on post-tests. In addition, it was also noticed that students' intellect correlated with higher physiological skin signal readings, with smarter students having a tendency to be more responsive (higher physiological skin signal measures). Ekman, Levenson *et al*. [177] investigated the relation between specific emotions and the autonomic nervous system (ANS) by positively and negatively stimulating study participants; then they measured their physiological signals, which included skin temperature, galvanic skin responses and changes in heart rate. They found that ANS responded differently to positive as well as to negative emotions; furthermore, they were also able to differentiate between negative emotions such as anger and sadness. Moreover, in an investigation by Kistler *et al*. [178], using fingertip temperature as an indicator of changes in the sympathetic nervous system due to stimulations, the study found that there was an immediate change in the measured fingertip skin temperature after using stimuli. This temperature change is due to the sudden increase or decrease in blood flow triggered by the sympathetic nervous system induced by the stimuli.

The study suggests that skin temperature is an effective yet simple way to evaluate changes in the sympathetic nervous system. Ward, Marsden *et al*. [179] reported the findings of their study of web pages with high quality in comparison to pages with lower quality having an identical content. They measured users' arousal while interacting with tested platforms by recording their physiological signals including skin conductivity; their results suggested that the mean value for recorded physiological data could be used as an indicator of web page usability. A study by Lin, Omata *et al*. [163] used physiological signals including galvanic skin response (GSR) in addition to conventional usability measurements, in assessing three video games with different conditions. The results showed a correlation between task performance and the obtained physiological results; their findings suggest that it is beneficial to use physiological measurements paired with conventional usability measurements. Baumgartner *et al*. [180] investigated the influence of music and images on emotion. The study measured physiological signals such as HVR, SC, skin temperature and breathing. The study findings suggest that music could improve emotional experience aroused by affective images. Grimshaw *et al*. [181] studied the sonic aspects of user experience with games, utilizing physiological signals including EDA; audio is commonly used in video games to guide users and respond to their moves. They found that the sound made games more enjoyable and players less stressful in comparison with no-sound playing conditions. Mandryk, Atkins *et al*. [182 and 183] used physiological signals, including GSR and HVR, in a study to model players' emotional states while using three different playing conditions in Interactive Playing Environments; the results show that modelled emotion has comparable trends to those emotions reported by the players such as excitement and amusement. Handri *et al*. [184] used galvanic skin response (GSR) to evaluate students' responses to interactive (virtual laboratory work) and non-interactive (video recorded lecture) e-learning course materials. The results obtained did not show any statistical significant difference between interactive and non-interactive materials and proposed that other physiological signals could be used in evaluation such as skin temperature combined with subjective data to evaluate e-learning course materials. Kätsyri *et al*. [185] used facial electromyography (EMG) to evaluate user aesthetic experience with the use of abstract images as a background for written messages which were displayed on small screens. The study found that arousing images could be employed to increase the user's skills in recalling written messages.

A study on 432 third-year medical students to investigate the effect of relaxation on students was carried out by Brennan, McGrady *et al*. [186], the results obtained showed that relaxation leads to statistically-significant increase in fingertip temperature, therefore temperature could be used as a physiological indicator of emotional experience. Sharma and Kapoor [107] attempted in an experiment to measure the physiological signals of emotions with a low-power embedded system; they used songs in Hindi and English to trigger participants' emotions. Their results show an increase in skin conductivity when participants felt sad, angry or anxious; however, it decreased with fear and there was no noticeable significant change with happiness.

Previous empirical studies [1, 2, 3 and 4] conducted by the Innovative Interactive Systems (IIS) research group used multimodal metaphors such as audio, video, earcons, auditory icons, synthesized as well as recorded speech and avatars in e-learning interfaces; like any conventional evaluation approach, the experimental platforms were tested for usability as this has always been perceived as an indicator for system quality. Based on the literature review, traditionally usability and its measurements of effectiveness, efficiency and user satisfaction has been used as a mean for insuring the success of any implemented computer system with human centred design. Usability evaluation only concentrate on the instrumental side and functionality of the system merely using an objective evaluation approach such as number of user errors ,calculating the times spent by the user to perform a particular task. Although usability has a subjective side within its user satisfaction measurements, however that only represent a tiny piece of the whole picture of user experience, also usability does not value the user affective state and its influence on user satisfaction and performance, which directly influence the measured system usability. In recent years, there have been growing interests in extending or replacing the concept of usability by user experience evaluation. The user experience concept surpasses the idea of simple user satisfaction adapted by usability; that is achieved through having a more holistic approach evaluating the pragmatic as well as hedonic qualities of the system and its overall attractiveness. However, this approach also does not address the objective side of the system-user relationship in its evaluation. User experience subjectively measures the pragmatic side of the system in terms effectiveness and efficiency, moreover it does neglects the evaluation of the user affective state and its impact on the user experience. One more move is in the field of human computer interaction is toward leaving behind the user subjective evaluation and calling for purely objective evaluation techniques, where the relation between the user and the system is judged and controlled by the system based on its interpretation of user's physiological signals (this concept has been gaining popularity as affective computing). This method lacks consistency a well as accuracy and disregards the subjective side of the user experience with the system, which is crucial for understanding the user perception of the system.

Using one or another of the aforementioned approaches does not provide a full realization of the system-user relationship especially in a system with a human-centred design. Another issues exist within the cognitive theory of multimedia learning CTML, although the theory represent a conceptual model for human computer interaction within the learning environment and how the visual and auditory channels work in delivering information between the user and the system. The theory does not explain how the system usability as well as the user experience and affective state influence the auditory and visual communication channels and their impact on the user cognition and learning performance within a multimodal interactive setting. Furthermore, the theory does not address other sensory channels and their contribution to the cognition process.

2.6 SUMMARY

A single evaluation approach on its own is not sufficient to guarantee success of a multimodal system; more attention must be paid to user experience and emotion, in order for the hedonic and pragmatic goals of the user to be met. Moreover, a holistic approach should be considered to overcome the gaps in the current evaluation practices and shortcomings of the CTML theory and achieving better understanding of relationship between the system and the user. We intend to adapt a triple evaluation approach, which employs objective as well as subjective techniques for the evaluation of system usability, in addition to measuring user experience and affective state. The proposed approach could help establish a better understanding of the system-user relation and help in delivering highly optimized and desirable multimodal solutions. Usability evaluation includes measuring effectiveness and efficiency and user satisfaction. Measuring user experience as well as affective state involves evaluating subjective feeling, felt emotion and physiological changes; therefore, the investigation will address those three dimensions using subjective and objective measurements. From the literature review and previous research, it appears that AttrakDiff instrument is suitable, well equipped, well supported, and capable of providing a sufficient evaluation of the subjective side of users and experiences. In addition, SAM Scale is useful as it can be used to evaluate the effect of the presented e-learning materials on the user's affective state. In order to observe expressed emotions and attitude towards the presented e-learning materials, FaceReader would make a good choice; however, due to financial restrictions, a more traditional observation method followed by unstructured interview to capture users' reflections on the presented material will be used. Biofeedback device will be used to capture changes in user skin conductivity and temperature while interacting with the system. Skin conductance is very suitable for detecting emotional arousal, because it is innervated only by the activation of SNS in contrast to other physiological signals, which are affected by more than one branch of the nervous system, which makes it difficult to distinguish whether an obtained signal is due to an emotional arousal or a biological process. Furthermore, skin conductance is easy to measure and interpret and is not affected by factors such as body movements. With regards to cost, measuring skin conductivity is cheap in comparison with other physiological measurements; moreover, it is convenient to use and does not require special training or an expensive setup to administer it. In terms of comfort, users feel comfortable with it. The QSensor biofeedback device from Affectiva [17 and 187] was selected. It was originally designed by the MIT lab[188], validated and backed by a great deal of research. The QSensor is a standalone device and provides a non-invasive measurement technology. Using the device involves neither wiring nor interruption while it is being used; the data obtained is saved on-board and can be retrieved for later use. In addition to skin conductivity, the QSensor measures users' skin temperature while they are interacting with the system. Skin temperature and conductivity can be used in conjunction to measure the affective state of the user.

CHAPTER 3

EXPERIMENTAL PHASE I: MULTIMODAL vs NON-MULTIMODAL

3.1 INTRODUCTION

This empirical investigation looks to test a multimodal e-learning condition in terms of user experience aspects (attractiveness and pragmatic and hedonic qualities), in addition to testing aspects of usability (effectiveness, efficiency and satisfaction) and users' affective state (valence, arousal and dominance). The multimodal e-learning condition combines a typical text with graphic metaphors and other metaphors such as speech, sounds and avatars with facial expressions in the delivery of learning information. The primary question is whether the inclusion of these metaphors can enhance user experience and the condition's usability and the secondary questions are:

What are the roles of these metaphors in any enhancement?

How do users feel about them?

What emotions do they trigger?

3.2 AIMS AND OBJECTIVES

The aim of this experiment is to investigate the impact of combining animation and speaking, facially-expressive avatars on the UX and usability of e-learning conditions. Furthermore, it is aimed at assessing the degree to which the inclusion of these multimodal metaphors could affect the user's engagement and learning performance as opposed to a typical graphical user interface.

In order to fulfil the aims of this experiment, the afterward objectives have to be met:

1. Formulate experimental hypotheses.
2. Create an experimental e-learning platform with two conditions, a Non-multimodal condition with text and graphics (NMMC), Multimodal condition (MMC) including speaking avatars, icons, recorded speech, auditory icons, text, graphics and classical music.
3. Adopt a within-subjects experimental design, where each user tests both conditions. This approach has some advantages as it eliminates errors due to individual difference and requires a smaller sample size.

4. Compare the obtained results from using both conditions to determine the roles of multimodal metaphors in any enhancement to the platform.
5. Evaluate the conditions' usability by:

   a. Measuring condition efficiency (time spent by users to complete the required tasks).
   b. Measuring condition effectiveness (by calculating users' correctly performed tasks), the measure can also be used to assess users' learning performance
   c. Measuring users' satisfaction.
   d. Measuring learning performance (learning time vs. scores).

6. Measure user experience (UX) with each tested condition using an AttrakDiff instrument [15] which employs a self-reporting technique to assist in performing the following:
   a. Measuring the pragmatic aspects of user experience.
   b. Measuring the hedonic aspects of the user experience.
   c. Measuring the condition's attractiveness.

7. Gain unbiased real-time insight into users' affective state by tracking changes in users' skin conductance while experiencing both conditions.

3.3 HYPOTHESES

Based on the type of conditions presented in this study, it is expected that the MMC will enjoy a higher rating and confidence by the users due to the addition of multimodal metaphors such as animation, earcons, auditory icons and speaking avatars. Based on the evaluation criteria and the measured variable; the following hypotheses have been formulated accordingly:

H1: The MMC is going to be more efficient than the NMMC with regards to the time spent by users to perform presented tasks.

H2: The MMC will be more efficient in performing tasks with higher complexity and cognition load.

H3: The MMC will accomplish higher effectiveness than the NMMC in correctness of answered tasks.

H4: The MMC will be more effective in performing tasks with higher complexity and cognition load.

H5: The MMC will be more satisfying for users than the NMMC.

H6: The users will have a better pragmatic experience with the MMC.

H7: The users will experience better identification quality with the MMC.

H8: The users will experience a better stimulating hedonic quality with the MMC.

H9: Overall, the MMC will provide users with a better hedonic user experience.

H10: The users will be more attracted to the MMC.

H11: The MMC will induce more positive valence.

H12: The users will have a better experience while using the MMC.

H13: The users will be more alert while using the MMC.

H14: The users will have better control over their interaction with the MMC.

H15: The MMC will generate higher electro-dermal activities.

H16: The MMC will positively influence users' affective state.

3.4 STUDY VARIABLES

The variables included in this study are grouped according to their type and listed in the following subsections (see to Table 3-1). In addition to the above-mentioned variables, other factors might have an effect on the results of the experiment and they need to be controlled:

3.4.1 Consistency

- All the testing equipment and machinery were synchronised.
- Desktop activities and the experimental sessions will be recorded.
- Users have to wear the skin conductivity sensor 10 minutes before they use the condition; this time lapse is necessary to establish their EDA baseline.
- Each user will experience both conditions.

- Experimental sessions with each user will be conducted individually, to eliminate any influence from other users.

Study Variables			
Variable Type			
Independent	**Dependent**		
NMMC (Control Condition) MMC (Experimental Condition)	1. Usability		1.1 Task completion time. 1.2 Correctness rate. 1.3 User Satisfaction. 1.4 Learning Performance.
	2. User experience		2.1 Pragmatic user experience. 2.2 Hedonic user experience. 2.3 Attractiveness.
	3. User affective state	3.1 SAM	3.1.1 Arousal. 3.1.2 Valence. 3.1.3 Dominance.
		3.2 Biofeedback	3.2.1 Skin conductance. 3.2.2 Skin temperature.

Table 3-1: Independent and dependent variables.

3.4.2 Procedure

Insuring and maintaining consistency in research is a crucial element for obtaining same procedure and steps will be pursued throughout the experiment; furthermore, all users will utilize the same tools and computer.

3.4.3 Ethical and Privacy Concerns

Users will be randomly selected from students around the university to participate in the experiment and all of them will be made aware of the following:

- Their privacy rights and identity will be protected.
- They have the right to terminate their involvement in the experiment at any time if they wish, without any questions asked.

3.4.4 Presented Tasks

Users will be required to perform a number of recall and recognition tasks and to insure they are balanced the following measurements were taken:

- All users will experience the same number of tasks in the test conditions.
- Users will be made aware of the task presented to them.
- Users will have a time range in which to complete tasks in hand; tasks completed correctly in time will be considered successful, otherwise they will be deemed unsuccessful.

3.4.5 Learning Topics

Although the experiment is using a within-subjects design, the learning topic presented in this experiment was selected to be a non-familiar topic to the users with the intention to minimize the impact, which their prior knowledge might have on their performance and to insure that the learning topic influence is controlled the following measurements were taken:

- Similar learning topics will be presented in both conditions with the same structure, sequence, and complexity.
- Platform and topic familiarity: All users are first time users and will be provided with the same presentation and training.

3.5 EXPERIMENT IMPLEMENTATION AND PROCEDURE

The conditions were designed to follow a linear sequence, starting by introducing the topics followed by the learning materials with gradually increasing complexity, then a quiz. The quiz contains two types of task, recall and recognition. The following software packages were used to build the experimental platform:

- Adobe Captivate [189], which is considered to be one of the best in the market for developing e-learning applications.
- Adobe Flash [190] was used for animation and illustration.
- CrazyTalk Animator Pro [191] was used for avatar animation.
- Adobe Photoshop was used for the editing of images [192].
- QuickScore Elite Level II MIDI Edition is a music composition software for Microsoft Windows [193] was used for the creation of earcons.
- Adobe Premiere Pro, a video editing software [194] was used.

The experiment involved 30 users. Over four-week period, participants tested the conditions on an individual basis. They were all university students, randomly picked and with no prior experience with the tested e-learning platform. The researcher who explained the experiment procedure, goals, users' rights, privacy and the confidentiality of their data were assured greeted each user. Furthermore, the researcher made it clear to them that the experiment was set to measure the condition performance and not to judge their own abilities or to judge them personally.

3.5.1 Experiment Execution

3.5.1.1 *First stage*

Step1. Users were requested to wear Affectiva QSensor [17], which is a non-intrusive wearable biofeedback sensor ,the QSensor measures two physiological bodily signals, the first one skin conductance which rises during the experience of emotional arousal due to excitement, anger and decreases in the course of experience of emotions such as relaxation. The second signal measure the QSensor is skin temperature, which is also affected by users' affective state.

Figure 3-1: Affectiva QSensor.

Users were asked to relax for ten minutes prior to using the condition in order to establish their SC baseline. As soon as the user wore the sensor, a green light started blinking indicating that the sensor was active and EDA measurements had started (Figure 3-1).

Step2. Users were invited to complete a pre-experimental survey to collect some information about them, such as educational background, age, gender, use of computers, use of the internet, previous experience with e-learning solutions and prior knowledge of the topic to be introduced by the e-learning platform (see Figure 3-2).

Step3. A short tutorial was given on how to use the platform and how to navigate from one stage to another.

Step4. Users were instructed to launch the condition and immediately press the sensor control button to mark their starting point of time using the platform; 50% of the users were presented with the MMC in the first experimental stage while the remaining 50% were presented with the NMMC. This procedure ensured that both conditions had an equal influence on the users and the obtained results (please refer to Table 3-2 for conditions rotation).

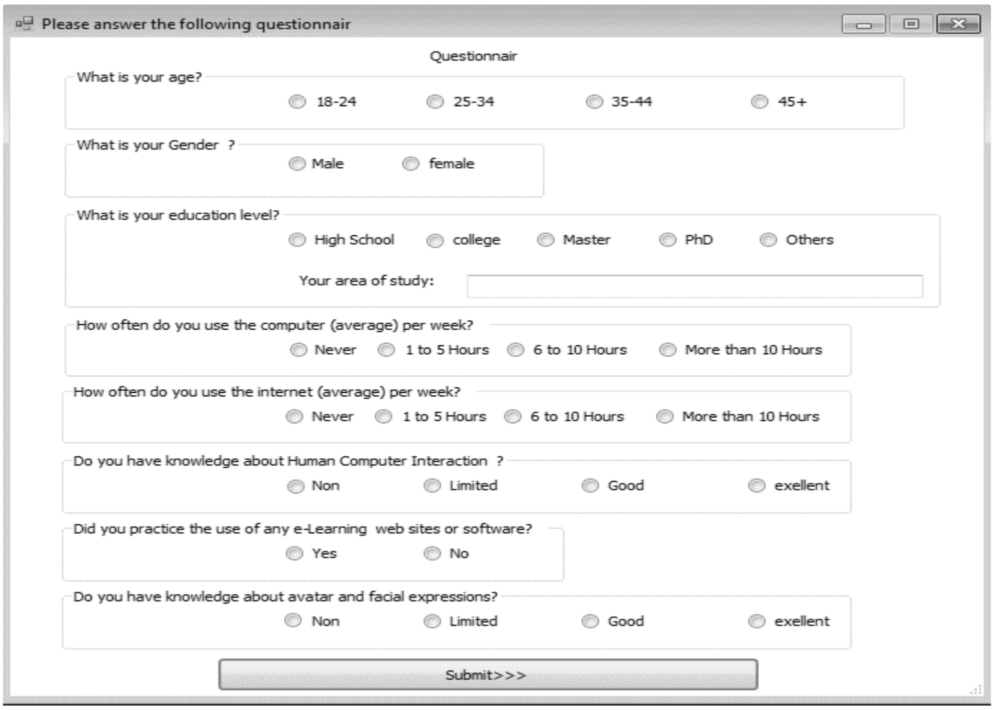


Figure 3-2: Pre-experimental survey.

User	**First stage**	**Second Stage**
1	MMC	NMMC
2	NMMC	MMC
…….	…….	…….
30	NMMC	MMC

Table 3-2: Rotation table of the tested conditions.

Within this step, users had to accomplish the following tasks:

- Learn about earthquakes as well as their causes and how they are initiated.
- Learn how the intensity of earthquakes is measured.
- Learn about the seismic waves caused by earthquakes.
- Learn how earthquakes are placed on maps.

Step 5: Users completed six tasks (divided equally between recall and recognition types) ranging from easy to moderate and then hard. They had to finish the task in hand before they could proceed to the next task.

Step 6: Upon completion of all tasks, the final score was displayed and a summary of the correctly performed tasks was presented. Then users were prompted to press the QSensor's control button to mark the end of the first experimental session.

Step 7: Users were instructed to complete an SAM's affective state questionnaire.

Step 8: Users were requested to use AttrakDiff instrument [157, 15 and 95] which was used to facilitate the evaluation of the conditions' hedonic as well as pragmatic user experiences and attractiveness. Upon completion, users were asked to relax for five minutes before continuing to the second phase of the experiment.

3.5.1.2 *Second Stage:*

During this stage, the user was introduced to the remaining conditions. Each user was requested to repeat all the steps performed in the first stage except for step 2, because all the necessary information had been provided in the previous stage. During both experimental stages, the researcher was watching closely, taking notes, and providing help to the user when needed. All the users' desktop activities during the experiment were recorded for later use in data collection.

3.5.2 **Users' Profile**

A total number of 30 users participated in the trial and successfully experienced both conditions. The users were made up of 67% males, 33% females, and various age groups. Looking into the educational background of our users we can see that 60% were undergraduates, 23% were Master's students and 17% were PhD students; furthermore they came from a variety of schools including Computing, Engineering, Business, Law, Media, Health, Art and Business Management (see Figure 3-3). The results show heavy use of the internet as well as computers with 97% using both for more than ten hours per week; in addition, all of the users were familiar with e-learning and had prior experience with e-learning websites and e-learning software. The majority of the users (67%) had no prior knowledge of the topic introduced by the e-learning platform; 23% had limited knowledge of the topic, followed by 10% who had good knowledge of the topic, while only 3% had excellent knowledge of the subject.

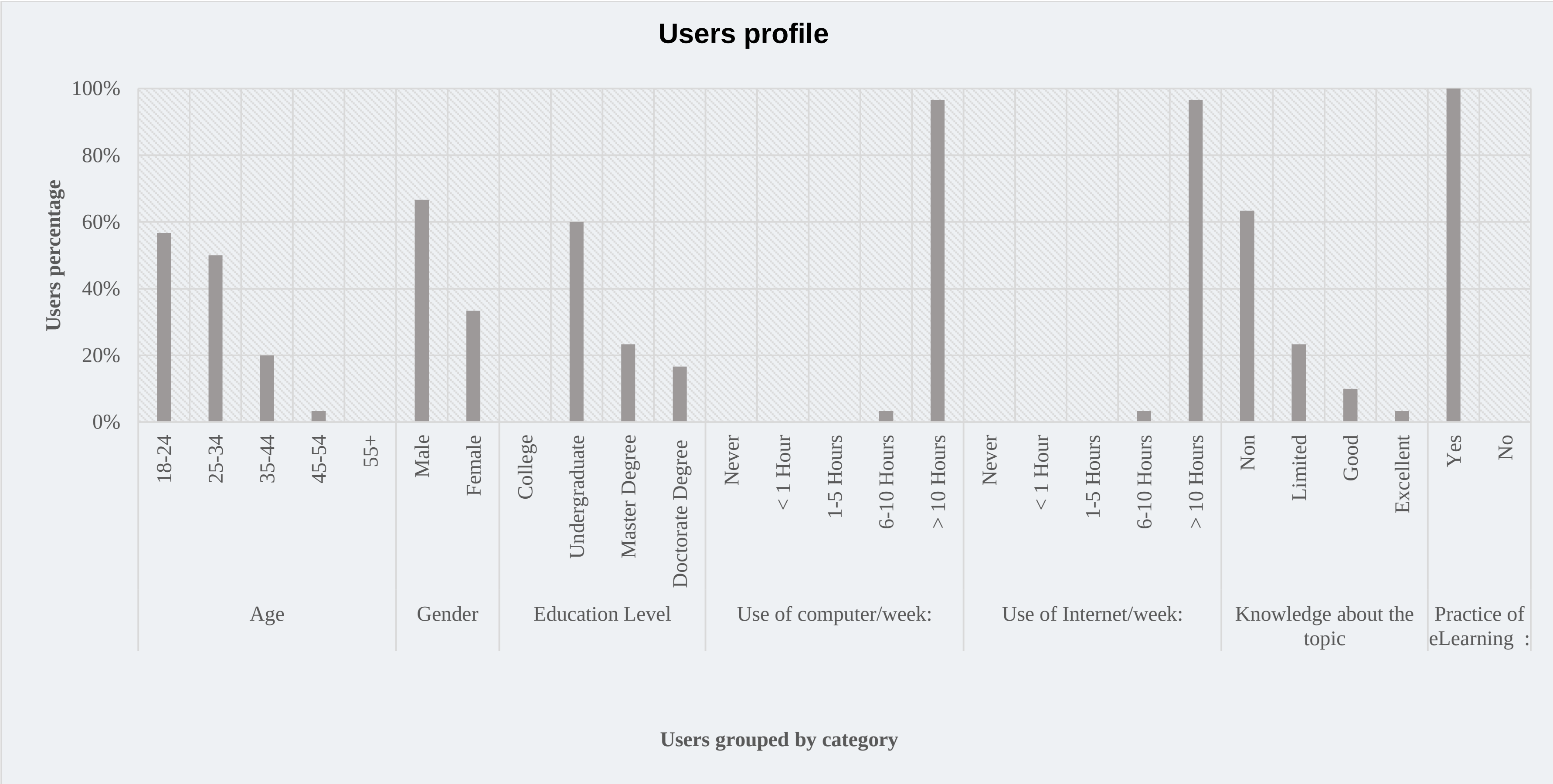


Figure 3-3: Users' profile.

3.6 RESULTS

3.6.1 Usability Evaluation Results

Usability is measured using its attributes including efficiency as well as effectiveness and user satisfaction. The collected data needs to be statistically examined. Selecting the proper statistical tests is critical for the validity of the evaluation outcomes, therefore, the data distribution was tested for normality using the Shapiro-Wilk test (recommended when the sample size is <50) [195]. In the cases where the data samples were found to be normally distributed, parametric statistical tests were used, such as a two related samples t-test. In other cases where the samples to be compared were abnormally distributed, a non-parametric tests were used depending on the type of data in use.

3.6.1.1 *Efficiency*

Efficiency is evaluated by measuring the resources consumed (time spent) by users to accomplish a particular task within the context of use [12]. In the experiment, users were required to perform twelve compulsory tasks varying in type and complexity. The tasks were equally distributed among the tested conditions according to their type and level of complexity.

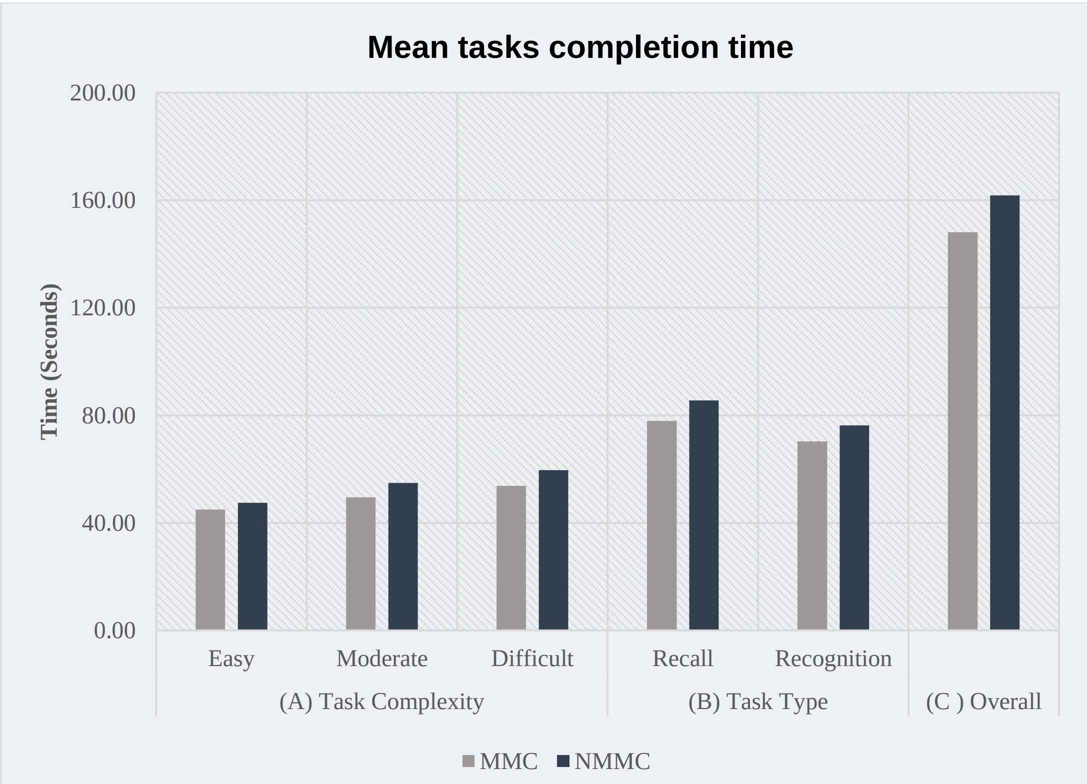


Figure 3-4: Task completion time.

Two types of task were used, recognition and recall tasks, with three levels of complexity, easy, moderate and difficult. The results show that the time spent in performing tasks increased gradually with the escalation of task complexity in both task types within both of the tested conditions (MMC, NMMC). The increase is depicted in Figure 3-4 (A, B, C). From the figure, it seems that in contrast to the MMC the NMMC required more time to accomplish the presented task in all types and at all levels of complexity.

All Tasks: Users needed a total of 3986.0 seconds to complete the NMMC task with a mean time of 132.86 seconds per user , while the overall time users needed to accomplish the tasks of the MMC was 3725 seconds with a mean value of 124.171 seconds; the result is displayed in Figure 3.4 (C). A two related samples t-test ($t(29) = -7.520$, $p < 0.05$) shows that the overall differences between the tested conditions in performing the presented tasks were statically significant. The test result confirms that, with the aid of the multimodal metaphors included in MMC, users were able to accomplish tasks faster than with NMMC.

Task Complexity: The tasks performed by users were categorized into three levels of complexity (easy, moderate, and difficult); the time spent by users to accomplish each level of complexity for both tested conditions was calculated and the results were compared to find any significant differences. Concerning the easy tasks, for the MMC the mean time was (18.46 seconds), while for the NMMC it was (19.25 seconds), the result ($Z= -2.582$, $p = .01$) of a Wilcoxon signed ranks test shows that the difference between users' performance at this level of complexity is statistically significant. With regards to moderate tasks, the mean time of the MMC was 20.73 seconds while the mean time for the NMMC was 22.89 seconds and in two related samples t-test ($t(29) = -6.840$, $p < 0.05$) the difference between the recorded time for users in both conditions at this complexity reached a significant level. Nevertheless for difficult tasks for the MMC as well as NMMC the mean time was (22.42 seconds) and (24.77 seconds) respectively; a two related sample t-test ($t(29)= -7.231$, $p < 0.05$) shows that the statistical difference in the accomplishment of tasks in both conditions was highly significant.

Task Type: Concerning the relation between task type and time spent by users, in the experiment users had to perform two types of task, recall and recognition (see Figure 3-5 C). The mean time for all recall tasks was 23.52 seconds and 21.95

seconds in the NMMC and the MMC respectively; a two related sample t-test (t (29)= -15.366, p <0.05) shows that the difference is highly significant. For all the recognition tasks, the mean time by users to accomplish the tasks was 20.67 seconds in the NMMC, while the mean time for the MMC was 19.34 seconds. A two related sample t-test (t (29) = -15.366, p <0.05) indicates that the difference in time needed by users to complete the recall tasks in both conditions was statistically significant. To investigate the observed differences further, it is worthwhile comparing task types in relation to their complexity levels, easy, moderate and difficult. See Figure 3-5 (A, B and C) and Table 3-3 for a summary of the comparison results.

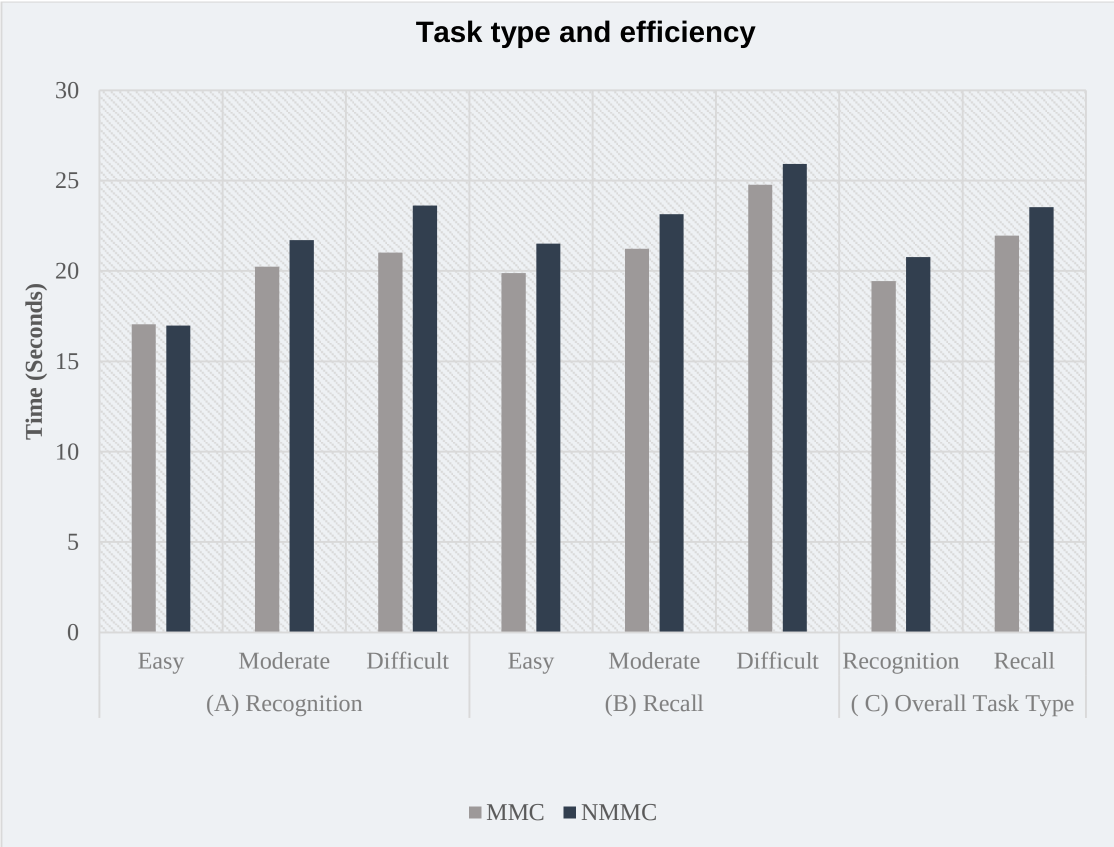


Figure 3-5: Task type and efficiency.

A two related samples t-test revealed a difference of high significance in easy (t (29) =-4.693, p <0.05) as well as moderate recall tasks (t (29) = -9.346, p <0.05) and in difficult recall tasks (t (29) = -5.401, p <0.05). While for recognition tasks a Wilcoxon signed rank test revealed that the difference in tasks with easy complexity (Z= -.257, p = .797) did not reach a statistical significance, whereas for the moderate complexity tasks (Z= -3.14, p = .02) the difference did reach a highly significant level.

Concerning difficult recognition tasks, a two related samples t-test confirmed that differences were of high statistical significance. Finally, it is safe to conclude that users, with the aid of the multimodal metaphors in the MMC (experimental condition), were able to accomplish tasks presented to them faster than they did in the NMMC (control condition) especially with tasks of higher complexity and cognitive load.

Significance Tests							
Task type	Conditions					Test	
	Complexity	MMC	Mean	NMMC	Mean	Test Result	Sig.
Recall	Easy	RCL1	19.88	RCL2	21.51	(t(29)=-4.693, p <0.05)	Yes
	Moderate	RCL3	21.22	RCL4	23.13	(t(29)=-9.346, p <0.05)	Yes
	Difficult	RCL5	23.135	RCL6	25.91	(t(29)= -5.401, p <0.05)	Yes
Recognition	Easy	RGN1	17.04	RGN2	16.97	(Z= -.257, p = .797)	No
	Moderate	RGN3	20.239	RGN4	21.69	(Z= -3.14, p = .02)	Yes
	Difficult	RGN5	24.765	RGN6	23.62	(t(29) =-5.979, p<0.05)	Yes

Table 3-3: Comparison of task performance time and complexity.

Each User: Regarding users' individual task performance time, the majority of users did better while using the MMC; however, there were three cases in which the user performed equally (U18 and U27) or faster (U11) when using the NMMC. Overall, users were able to perform faster in the MMC with a mean of 20.7 seconds compared to 22.15 seconds while using the NMMC (refer to Figure 3-6 for illustration).

3.6.1.2 *Effectiveness*

Users achieve evaluating the effectiveness of the experienced conditions through calculating the percentage of correctly performed tasks. The users were asked to perform two types of task (recall and recognition); each task consisted of six questions, which varied according to their levels of complexity (easy, moderate and difficult).

Figure 3-7 shows the percentage of correctly performed tasks grouped by task complexity and task and condition type. It is noticeable that users achieved better results using the MMC in all task types and at all complexity levels. The raw data of users' answers is available in Appendix B.

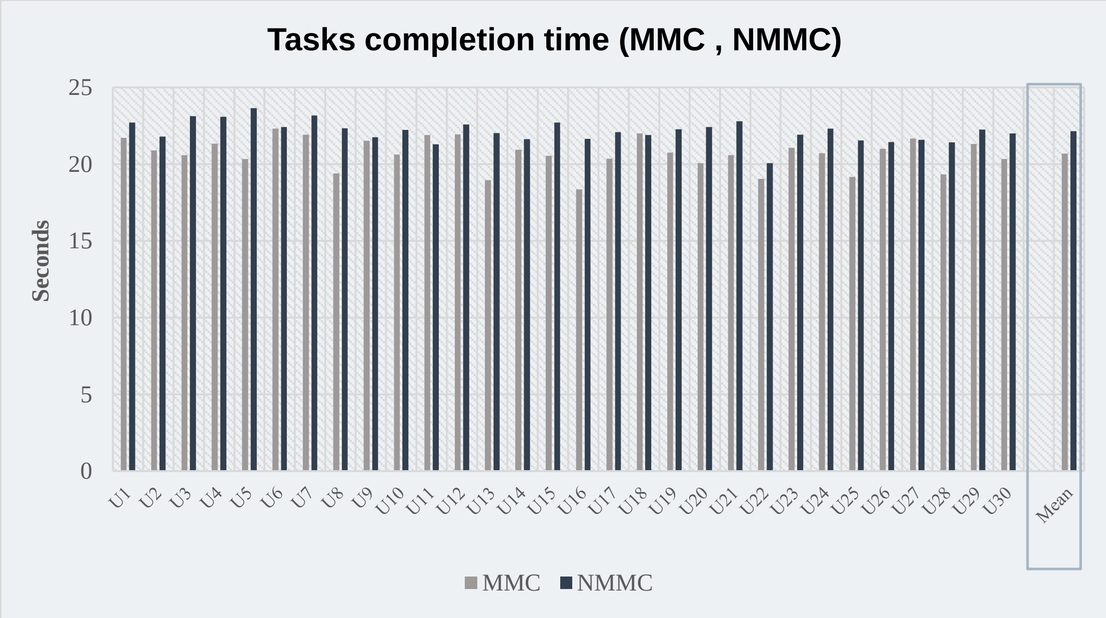


Figure 3-6: Tasks completion time (M).

3.6.1.3 *Conditions and Tasks*

In this experiment, 30 users were asked to answer six questions for each tested condition. The total number of questions was 360 (6 questions *2 conditions*30 users). Figure 3-7 (C) shows that the overall correctly performed tasks in the MMC totalled 79%, which is 21% greater than the result scored using the NMMC. The total number of correctly answered questions for the MMC was 152 compared to only 111 in the NMMC; a Wilcoxon signed ranks test was used to compare the obtained results for both conditions. The test result (Z=-3.971, p =.000) clearly indicates that, there is a significant variance between the tested conditions in terms of correctly performed tasks; it appears that the MMC helped users to achieve better results in performing the presented tasks. These improvements are attributed to the use of a combination of multimodal metaphors (auditory icons, classical music, earcons, expressive avatars and recorded speech).

3.6.1.4 *Task complexity*

In Figure 3-7 (A) correctly performed tasks are displayed according to their complexity levels. There are three levels of complexity, easy, moderate and difficult. Each level includes 60 questions, with a total number of 180 questions per condition. It is evident that the MMC condition generally outperformed the NMMC at all levels of complexity; however, this was more obvious within moderate and difficult tasks. In performing easy tasks, users scored 6.7% higher in the MMC; concerning moderate tasks, the MMC exceeded the NMMC in the number of correctly performed tasks by 26.7%, whereas the highest difference was obtained when performing, difficult tasks in which the NMMC scored 35% lower than the MMC in the number of correctly accomplished tasks.

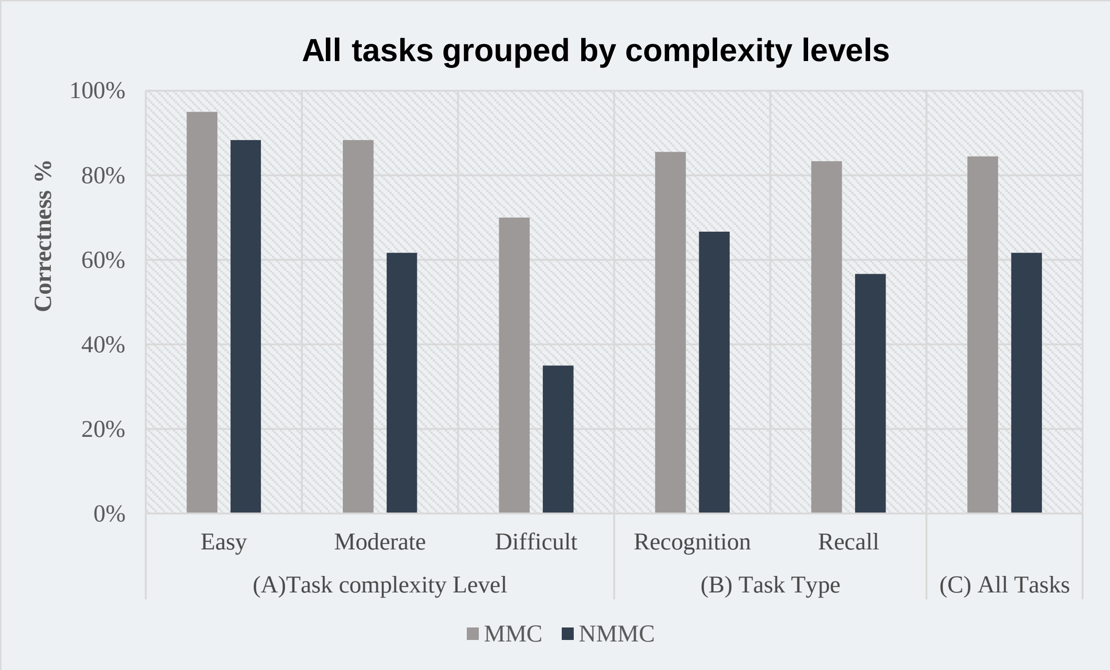


Figure 3-7: Tasks effectiveness (grouped by their complexity levels).

A Wilcoxon ranked test shows that the difference between the two conditions in performing easy tasks (Z= -1.414, p =.157) did not reach a significant statistical level. However, the differences were statistically significant in performing moderate (Z= -2.909, p =.04) and difficult tasks (Z= -3.275, p =.001). Overall, it can be concluded that, while both conditions were comparably effective in performing easy tasks, the MMC was more effective in accomplishing tasks with higher complexity levels and the interaction metaphors significantly increased the effectiveness of the MMC "experimental condition".

3.6.1.5 *Task type*

Figure 3-8 (C) shows ,that the MMC outperformed the NMMC in the number of correctly answered questions in both types of task; furthermore the figure shows that both conditions performed better in recognition tasks than in recall tasks (A and B) illustrates the percentage of correctly performed recognition and recall tasks for the MMC and the NMMC. Each task consisted of 90 questions per condition, with a total number of 360 questions for the whole experiment. It is clear in Figure 3-8 (C) that the MMC outperformed the NMMC in the number of correctly answered questions in both types of task; furthermore, the figure shows that both conditions performed better in recognition tasks than in recall tasks. The correctness rate acquired by users with the MMC was 25.86% higher in performing recall tasks and 22.82% higher in performing recognition tasks.

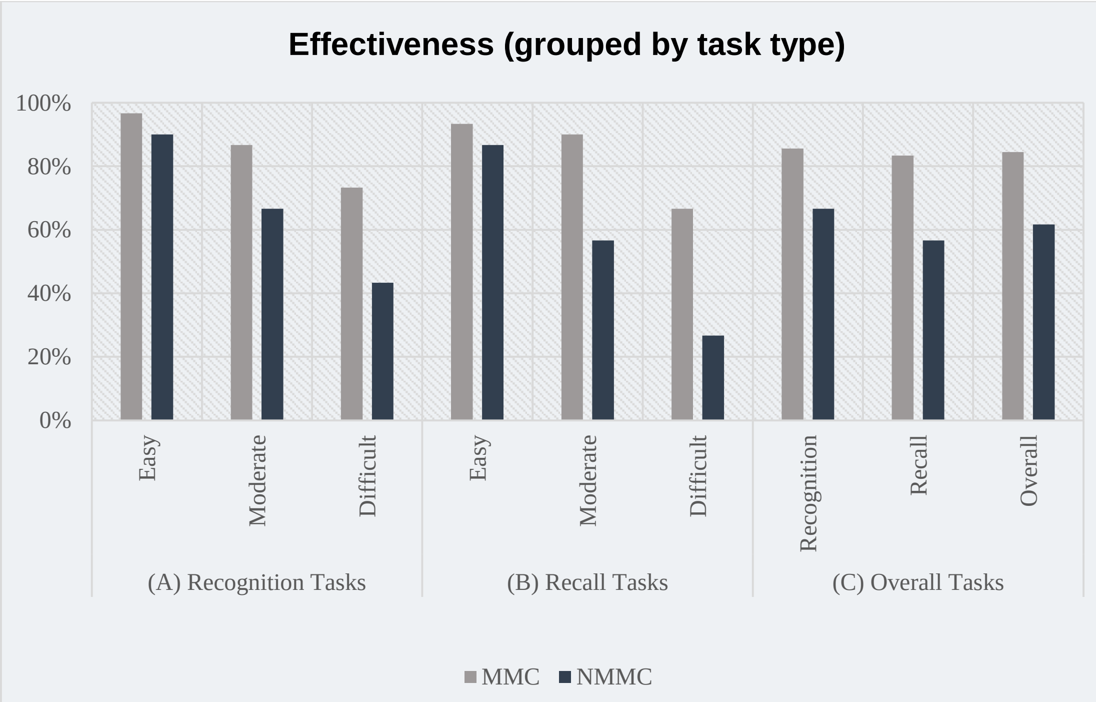


Figure 3-8: Effectiveness and task type.

A Wilcoxon ranked test shows statistically significant variances between the MMC and the NMMC in performing recall (Z=-3.404, p =.001) and recognition (Z=-3.206, p =.001) tasks. Looking deeper into the types of task with reference to their complexity levels in the tested conditions, the McNemar test shows that there are significant differences within difficult recall (p =.013) and recognition (p =.035) tasks. Furthermore, in the moderate recall tasks the difference (p =.013) is highly significant; however for the moderate recognition tasks the difference (p =.146) did not reach a statistical significance. In addition, for easy recall tasks (p =.688) and recognition tasks (p =. 625) the differences did not reach a significant level. Overall, it is obvious that utilizing multimodal metaphors made the MMC more efficient, particularly in performing recall and recognition tasks with higher complexity levels.

3.6.1.6 *Satisfaction*

To measure the satisfaction aspects of the test conditions' usability, a modified version of the System Usability Scale (SUS) [196 and 197] was used. The scale includes the 10 original SUS statements, which are divided into five positive, and five negative to provide users with a balanced view [198]. The modification is made to the SUS's rating scale, the "undecided" choice was eliminated and the rating gauge was downsized into 4 points only (1="Strongly disagree", 2="Disagree", 3="Agree" and 4="Strongly Agree"). The alteration was made to eliminate the issue faced in previous studies by our research group whereby a high percentage of the participants were undecided on their liking of the tested systems. The alteration is expected to push the users to express their true opinion of the system. To investigate user satisfaction with the aspects of the tested conditions further, eight additional statements were included and tailored to each condition.

SUS Results: The final SUS scores for both conditions are displayed in Figure 3.9; the negative statements were reversely coded and the final scores were calculated according to the SUS-recommended calculation method [196]. The outcome of the MMC-SUS test was ($M=83.91$, $SD=10.\ 28$), whereas for NMMC-SUS it was $M=64.26$, $SD=10.99$; this is just below the average acceptable value of 67%. A paired samples t-test (at $(29) = 6.67$, $p <0.05$) shows a significant difference between the two conditions; it is safe to conclude that the MMC scored better and seems to be more usable. The majority of the users found learning how to operate the system and using it to be an easy process in both conditions. In terms of user confidence in the system, users showed greater confidence in the MMC (97%) in comparison with the NMMC (50%). In terms of system simplicity, integrity, consistency and need of technical support, both conditions were highly regarded by the users. On asking users whether they would prefer to use the system often, the MMC achieved a higher score with 93% in contrast with the NMMC where only 39% showed an interest. The outcome of the MMC-SUS test was ($M=83.91$, $SD=10.\ 28$), whereas for NMMC-SUS it was ($M=64.26$, $SD=10.99$); this is just below the average acceptable value of 67%. A paired samples t-test (at $(29) = 6.67$, $p <0.05$) shows a significant difference between the two conditions; it is safe to conclude that the MMC scored better and seems to be more usable. The majority of the users found learning how to operate the system and using it to be an easy process in both conditions. In terms of user confidence in the system, users showed greater confidence in the MMC (97%) in comparison with the

NMMC (50%). In terms of system simplicity, integrity, consistency and need of technical support, both conditions were highly regarded by the users. On asking users whether they would like to use the system more often, the MMC achieved a higher score with 93% in contrast with the NMMC where only 39% showed an interest.

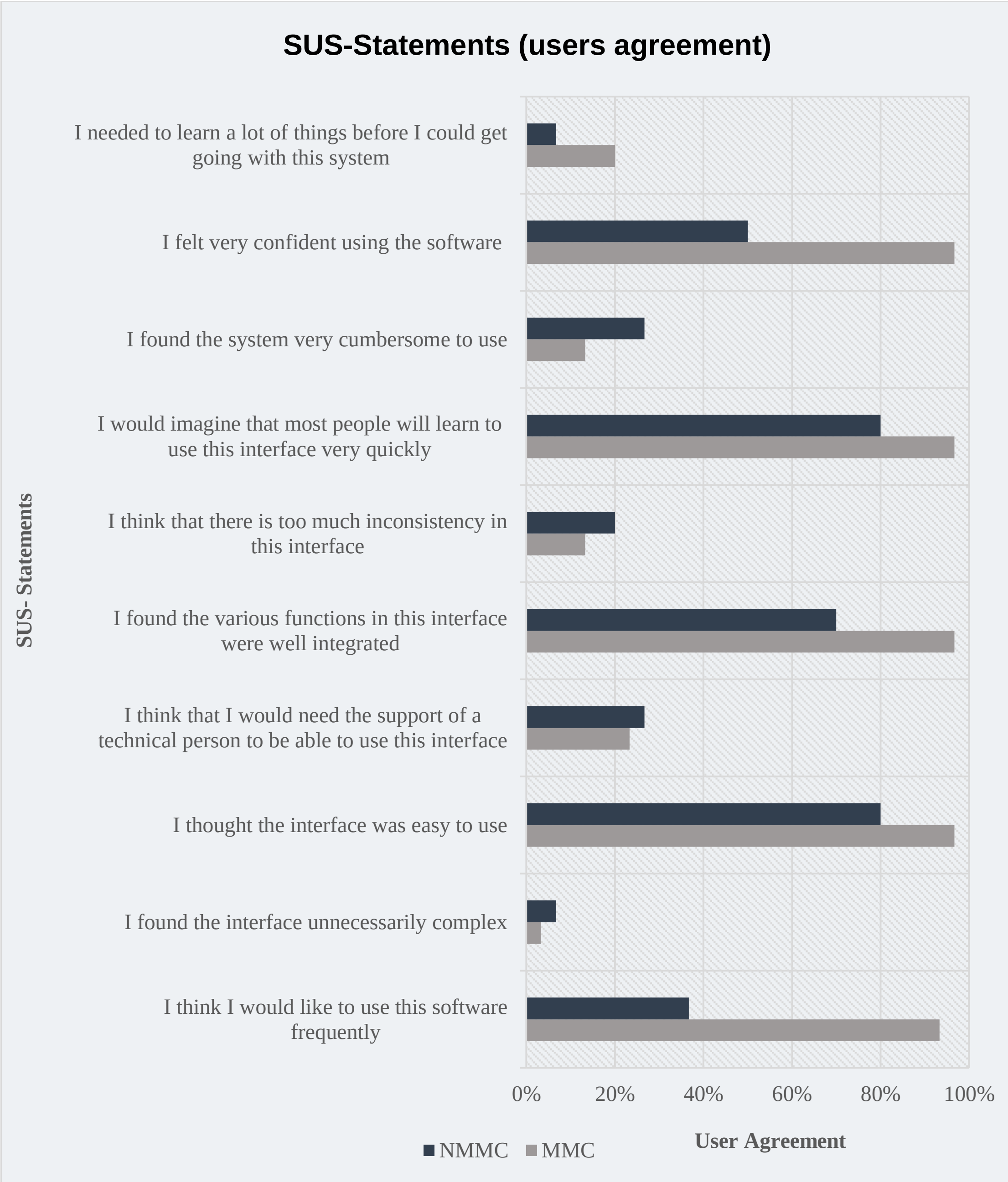


Figure 3-9: Users agreements with SUS statements.

Additional Satisfaction Statements: All users favoured the MMC in their rating (please refer to Figure 3-10 for more user ratings and Figure 3-11 for user agreements with the additional statements for both conditions).

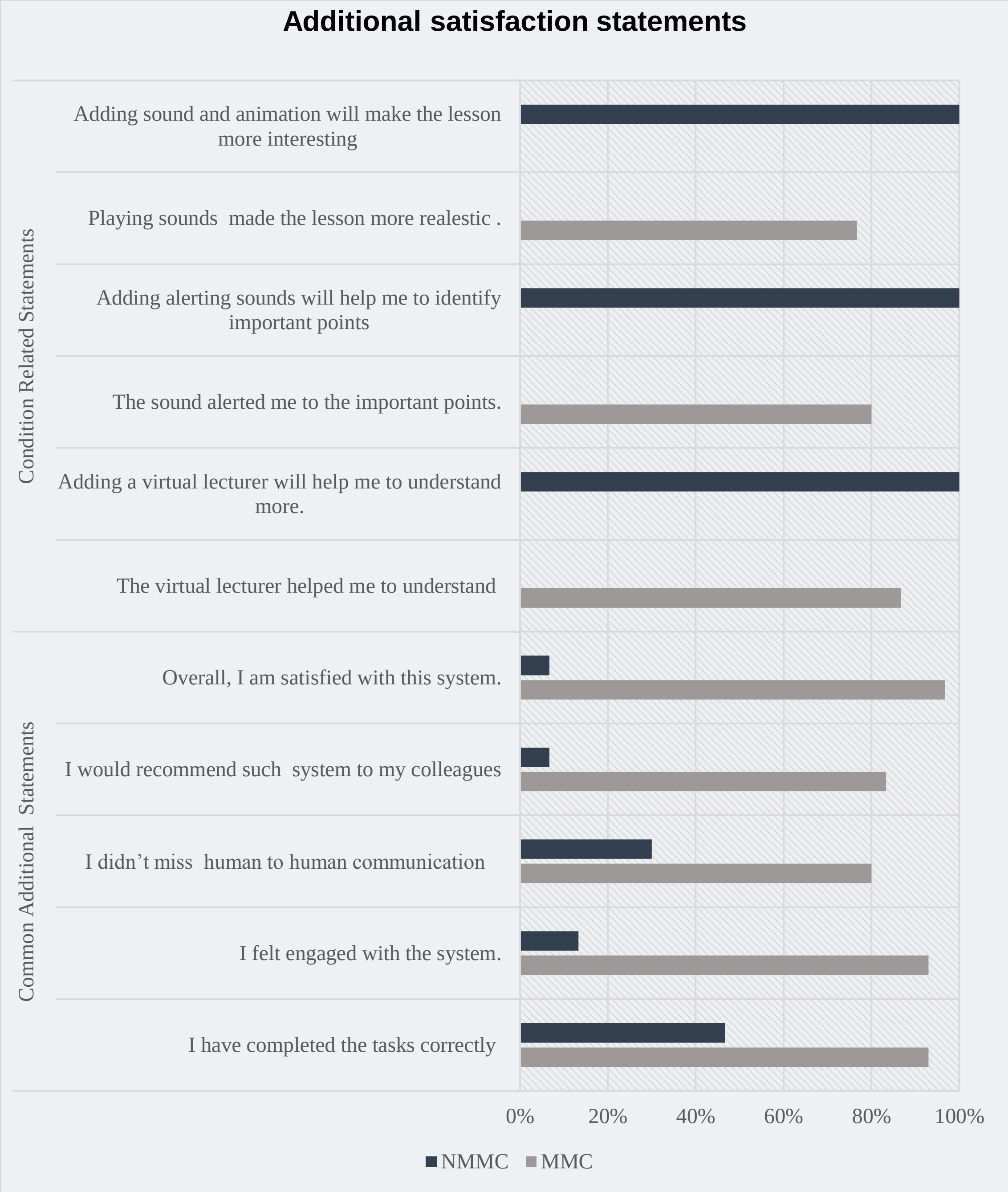


Figure 3-10: User ratings for each additional satisfaction statement.

The additional statements can be grouped into two categories; four of them are conditions-related statements while the other four are common statements for both conditions (please refer to Figure 3-10 where users’ overall rating is displayed for each statement). Users showed greater confidence in their skills and ability to accomplish the presented tasks in the MMC (93%) in comparison with only 47% who felt the same while using the NMMC. In assessing user’s engagement with the system, only 13% of the users thought they were engaged with the NMMC in contrast with 97% in the MMC; similarly, 70% felt they missed human-to-human communication while interacting with the NMMC condition, as opposed to only 20% for the MMC. It seems that users assimilated and connected very well with the MMC.

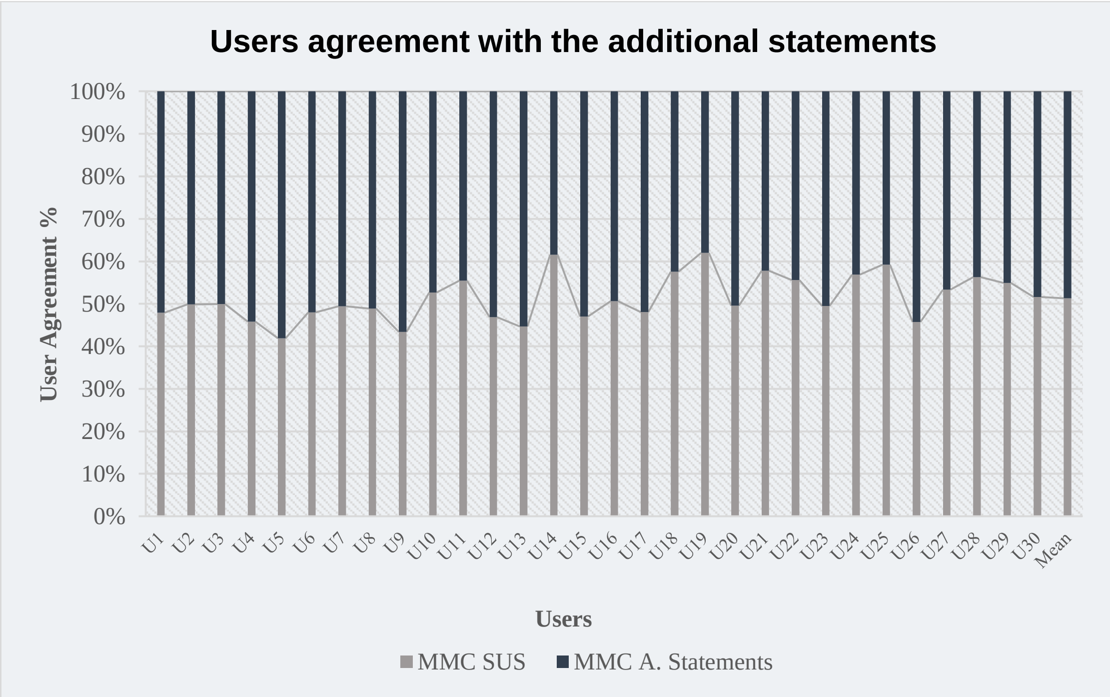


Figure 3-11: Users’ agreements with the additional statements.

When asked to state to what extent they were satisfied and whether they would recommend such system to a friend; 97% of the users had a negative view of the NMMC; in contrast, the user rating for the MMC condition was highly positive. For the NMMC all the participants thought that adding a virtual lecturer, sound, and animations would have improved their learning and engagement with the system. Concerning the MMC, 87% of the users stated that the presence of a virtual lecturer was positive to their engagement and learning outcome; more than 77% thought adding sounds to the MMC made the lesson more interactive and improved their alertness to the important topics. Generally, users expressed more interest and satisfaction in the MMC. In the SUS evaluation, the MMC scored 21.65% higher than the MMC; similarly, for the additional satisfaction statements, users' views toward the MMC were almost 20% higher. There is strong evidence that the MMC was more satisfying to the users (please refer to Figure 3-12, where the overall scores are illustrated).

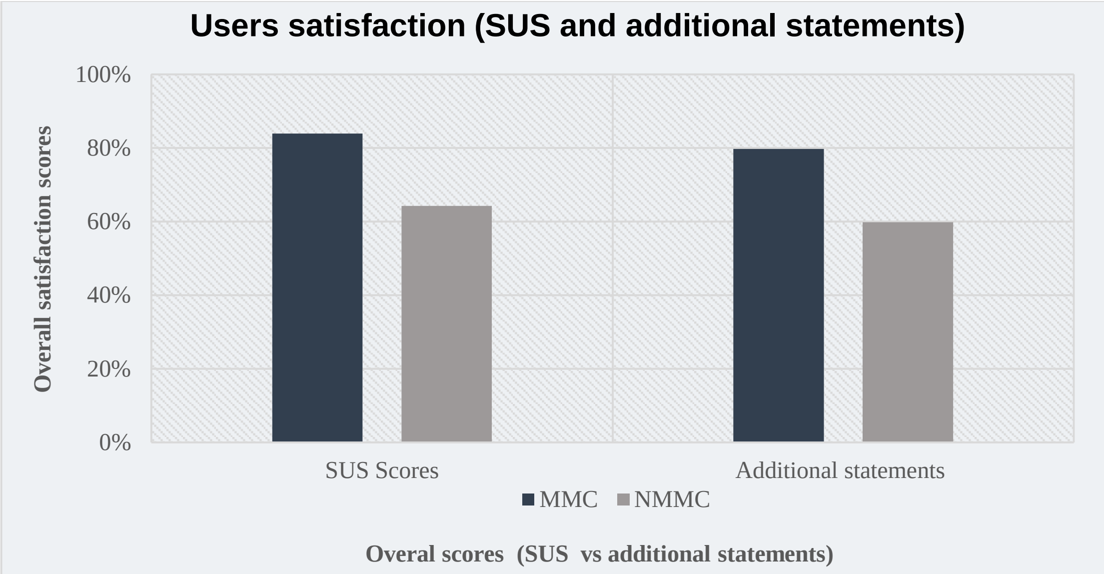


Figure 3-12: Overall view of users' satisfaction.

3.6.1.7 *Learning Performance*

Learning is a cognitive process, which takes place within a limited time frame, and its outcome is measured by learners' ability to grasp the presented learning topic; therefore, learners' performance enhancements in this study were valued within those two dimensions. Figure 3-13 presents the actual learning time and users' scores in performing the required tasks. To clarify the relationship between learning time and learning outcomes a scatterplot was created (Figure 3-14). The figure shows that the MMC (score=84.33%, learning time=281.61s) outperformed the NMMC (score=61.7%, learning time=287.43s) in learning performance.

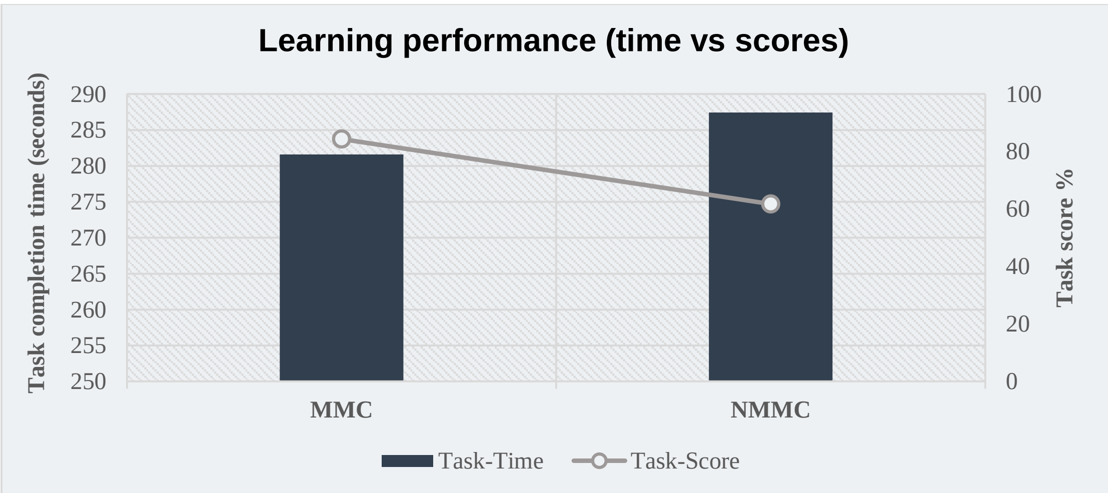


Figure 3-13: Learning performance comparison (MMC vs NMMC).

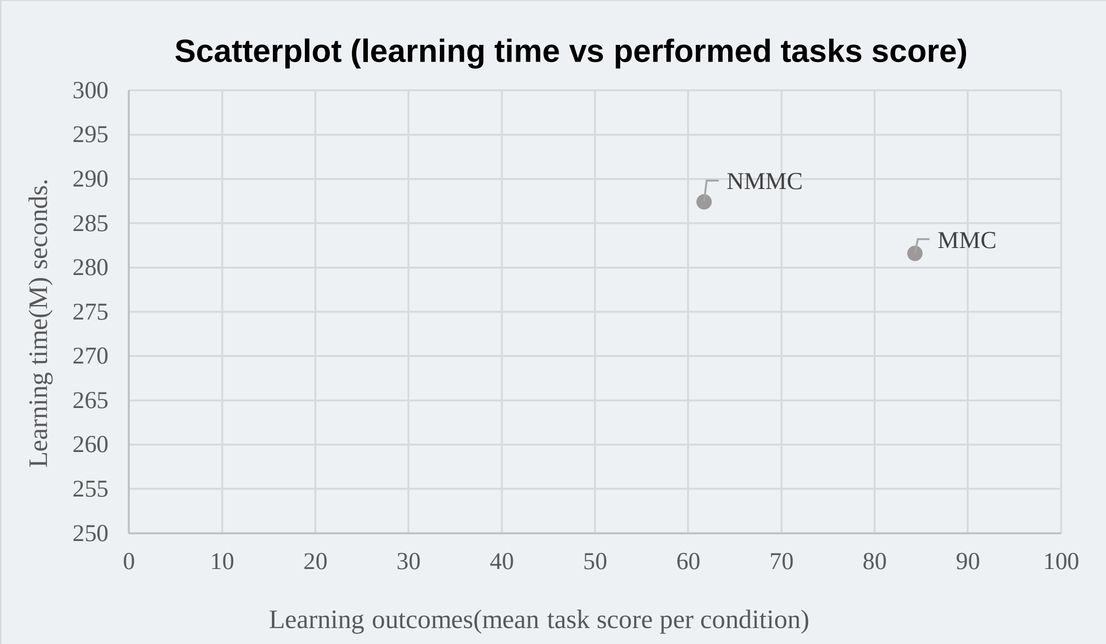


Figure 3-14: Learning performance scatterplot (learning time vs scores).

For the association between learning time and learning outcomes concerning learning performance as well as condition type, the following two hypotheses were formulated:

- H0: there is no correlation between learning time and outcomes.
- HA: there is a correlation between learning time and outcomes.

Pearson Correlation Analysis (SPSS2013) was used to obtain the correlation coefficient and significance level and the outcomes are shown in Table 3-4.

Correlations			
		MMC_Score	MMC_Time
MMC_Score	Pearson Correlation	1	.182
	Sig.(2-tailed)		.336
	N	30	30
MMC_Time	Pearson Correlation	.182	1
	Sig.(2-tailed)	.336	
	N	30	30

Table 3-4: MMC Pearson correlation analysis result.

Correlations			
		NMMC_Score	NMMC_Time
NMMC_Score	Pearson Correlation	1	.070
	Sig. (2-tailed)		.711
	N	30	30
NMMC_Time	Pearson Correlation	.070	1
	Sig. (2-tailed)	.711	
	N	30	30

Table 3-5: NMMC correlation analysis result.

Concerning the NMMC The result are listed in Table 3-5. There was more significance in the correlation between learning time and scores in learning performance ($r = 0.711$, $\rho=.070$) within the MMC. The results show that increase in learning time had a positive impact on users' scores. Therefore, the improvements observed within the MMC can only be attributed to the use of the communication metaphors. Overall, it is safe to conclude that the inclusion of communication metaphors could improve users' learning performance by lowering learning time and increasing information retention.

3.6.2 **User Experience Evaluation**

AttrakDiff instrument was used to evaluate user experiences in terms of pragmatic as well as hedonic qualities and attractiveness. The dimensions of user experience are independent and equal in their contribution to the overall score of both conditions.

3.6.2.1 ***AttrakDiff Results for the MMC***

The multimodal condition was ranked as "highly desired". Pragmatic quality is evidently the classification. The users were aided by the condition, the value of pragmatic quality reached the above-moderate values and there is an opportunity for enhancement. Concerning the hedonic quality, the character classification clearly fit the condition. It is strongly hedonic. The users identified themselves very well with the condition and are encouraged and motivated by it.

The users were in agreement in their rating of both dimensions (see Figure 3-15 for details). The NMMC was rated as "neutral". The users were assisted by it, nonetheless the score of pragmatic quality just reached the ordinary values and consequently there is certain room for enhancement.

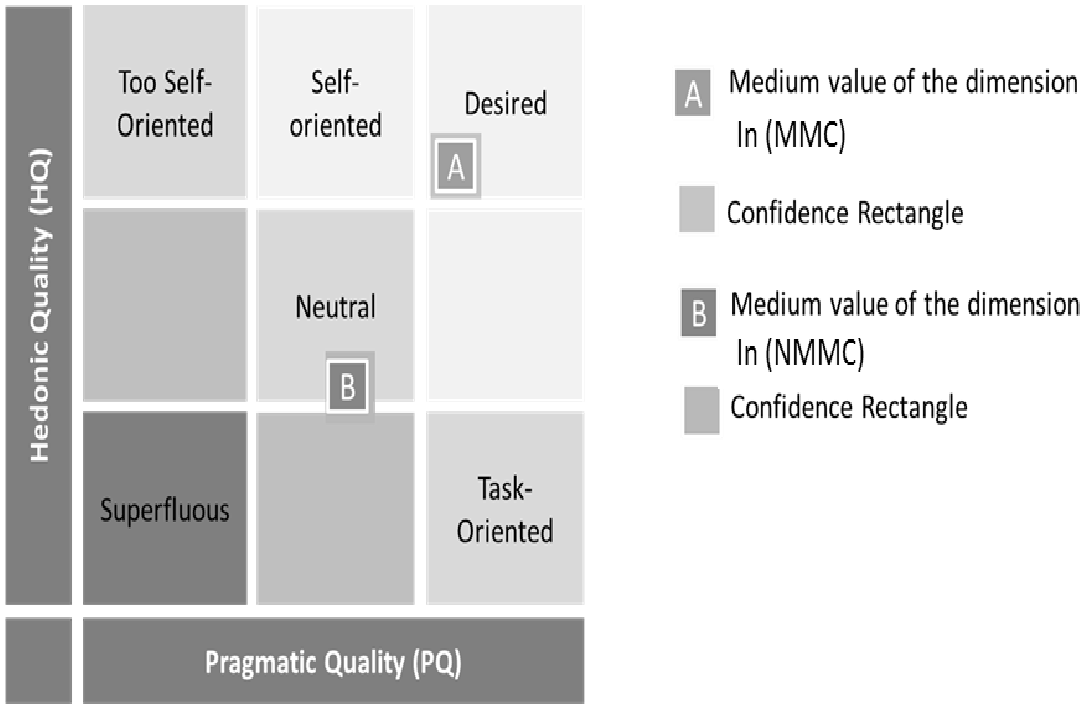


Figure 3-15: AttrakDiff results for the MMC and the NMMC.

3.6.2.2 *AttrakDiff Scale Results for the NMMC*

With regards to hedonic quality, the term description plainly does not fit the evaluation criteria for the reason that, the confidence interval stumbles over another zone. The hedonic value is only reached the ordinary levels and concerning hedonic quality, there is certainly a room for expansion. The MMC performs better than the NMMC in terms of pragmatic quality and hedonic quality. The difference in hedonic quality between multimodal and non-multimodal is statistically significant. The difference in pragmatic quality between multimodal and non-multimodal is also statistically significant (see Figure 3-15 for details).

3.6.2.3 *Figure of Average Values*

The overall values of AttrakDiff for the assessed conditions are depicted in Figure 3-16. Within this illustration, hedonic quality differentiates between the sides of stimulation and users' identification with the conditions. Concerning the MMC, in terms of pragmatic quality the condition is placed in the above-average region and meets ordinary standards. It is essential to improve the condition in this area even more.

Concerning HQ-I, the condition is placed above the moderated region. It motivates users, awakens their interest and encourages them.

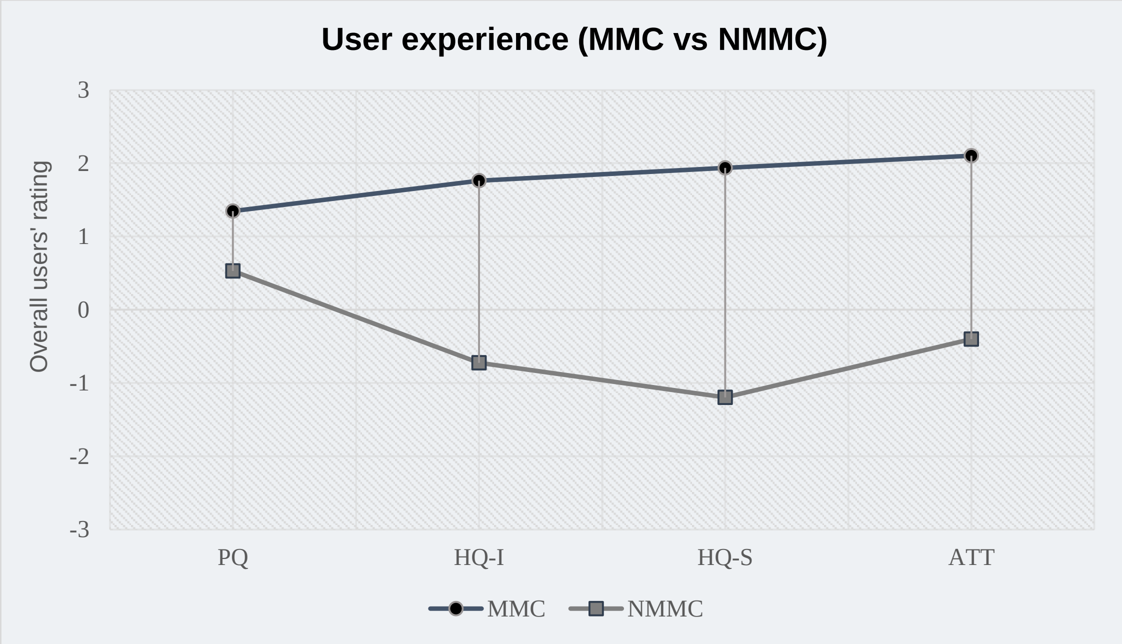


Figure 3-16: AttrakDiff average values.

It motivates users, stimulates and interests them. In terms of aspects of stimulation, the condition is also classified as ideal. The condition’s attractiveness rating is positioned above the ordinary region. The general impression of the condition is highly attractive. For the NMMC, in terms of pragmatic quality the condition is placed in the regular region. It just about come across regular standards and aiming at improving is needed. With respect to hedonic HQ-I, the condition is placed in the ordinary region. It meets ordinary standard, so in order to bind the user to the condition, aiming for improvements is necessary. With regard to HQ-S, the condition is placed in the below-average region. Furthermore, the condition did not have an inspiring influence on others. Inadequate inspiration in response, results in lack of motivation by the use of the condition and if a system of similar pragmatic quality were to be offered, users would happily switch to it; the condition is in serious need of revision.

The condition's attractiveness score is placed below the average region; the general impression of the conditions is slightly unappealing. For an overall outlook on users’ experience with the tested condition, see Figure 3-17 where paired users rating is depicted and compared.

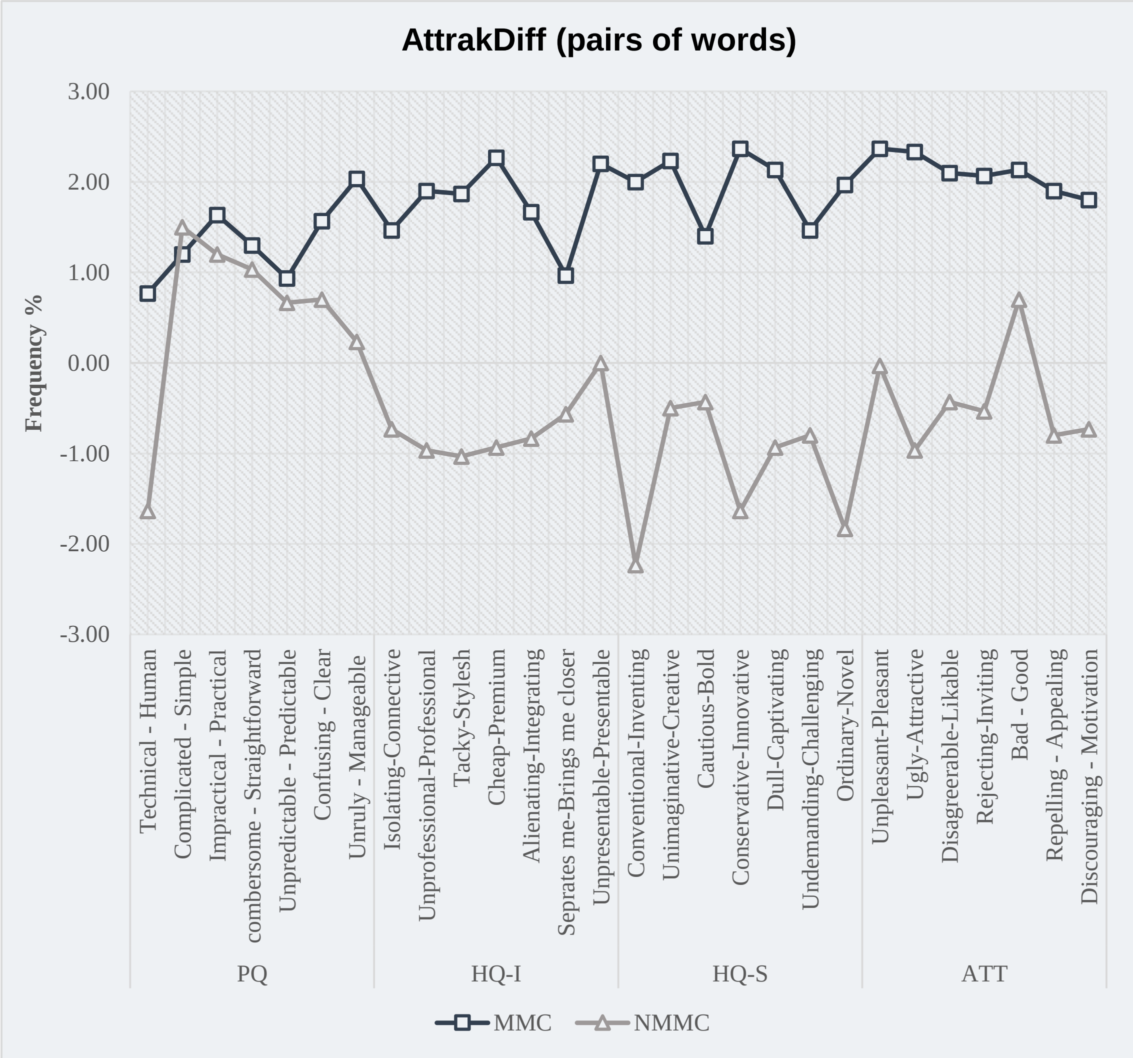


Figure 3-17: Word-pairs rating diagram.

3.6.3 User Affective State Evaluation

3.6.3.1 *Biofeedback*

Skin Conductance: Users' skin conductivity was monitored and the data was collected using Affectiva QSensor in the two experimental conditions. For each user, the mean value of SC, SKT was calculated. In the NMMC condition, the overall mean value for SC was 2.197μ $semeines$, SD=2.338, whereas in the MMC it was 2.755μ $semeines$, 2.773 in the NMMC. The highest obtained mean value of SC for a user was 11.79μ $semeines$ and the lowest was 0.1236 μ $semeines$, while for the MMC the results were 13.183μ $semeines$ and 0.168μ $semeines$ respectively (see Figure 3-18 for illustration and for an overall comparison).

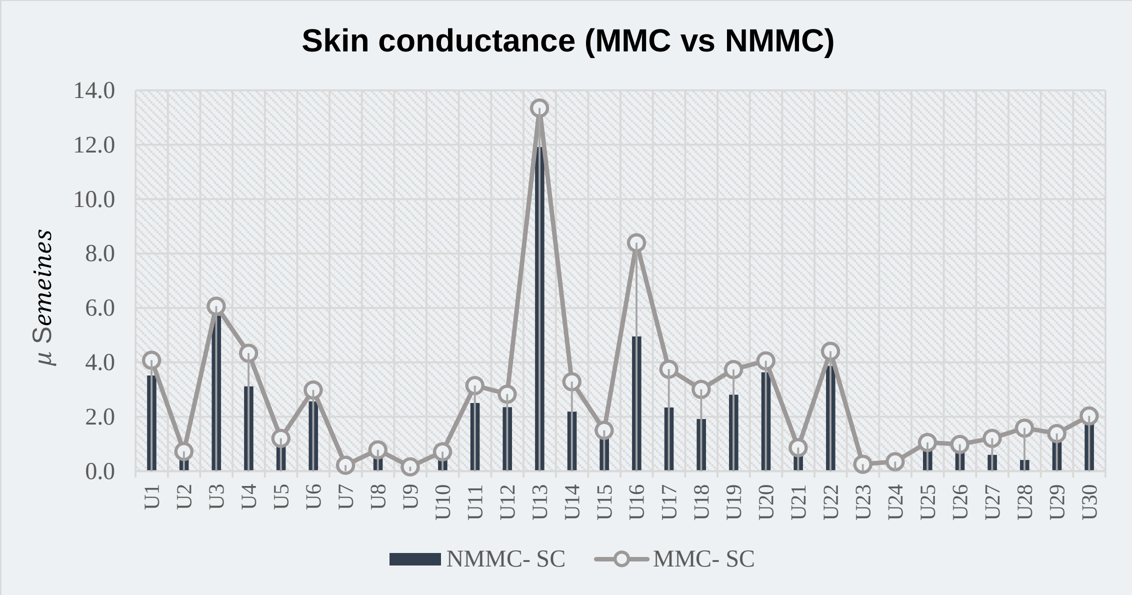


Figure 3-18: Mean skin conductance measured in Microsiemens (μS).

The results obtained show an increased level of users' skin conductance while using the MMC in comparison with the NMMC; a two related samples Wilcoxon signed ranks test shows the difference (Z= -3.363 , P=.001) is highly significant, the result shows clearly that the users were more excited while interacting with the MMC. Skin conductance provides an indication of the overall changes in electro-dermal activities, which mainly result from emotionally elicited bodily arousals.

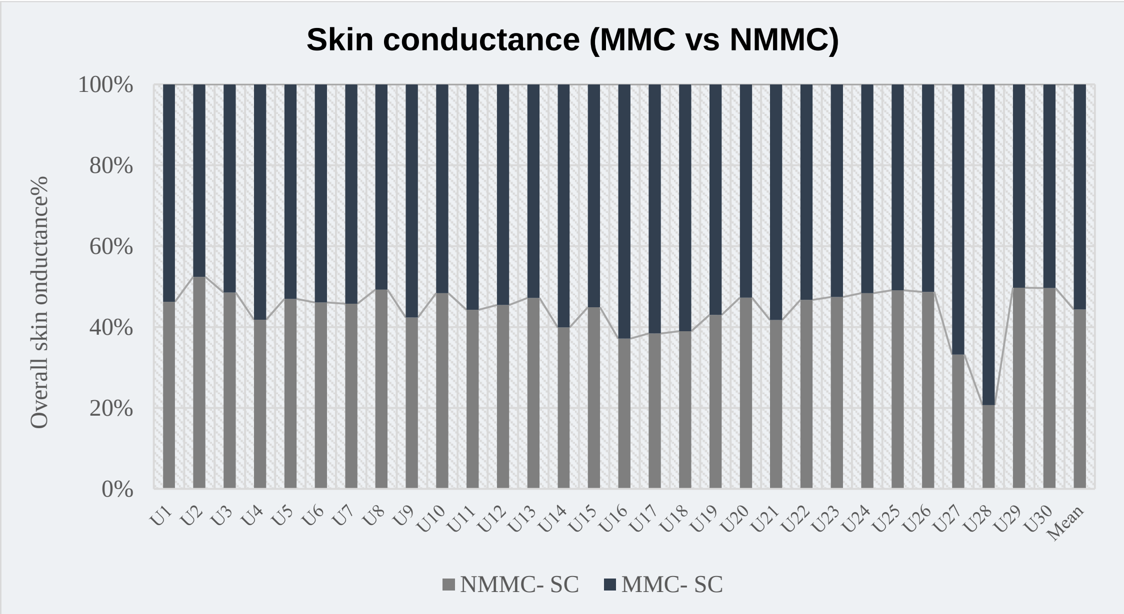


Figure 3-19: Overall skin conductance comparison.

Skin Temperature: In addition to skin conductance, the biofeedback device also records its electrode's temperature, which is reflective of changes in skin temperature. The data was collected for all users while they were experimenting with the MMC and NMMC. Skin temperature is to some extent similar to the tonic skin conductance data as the changes are small and gradual over time (the results are displayed in Figure 3-20). The mean skin temperature of users of the MMC was 30.91 C° while for the NMMC it was 31.05 C°. Overall, it seems that using the NMMC slightly increased users skin temperature. A two related samples Wilcoxon ranked test (Z= -2.92, P=.003) shows that the difference between users' skin temperature in both conditions reached a statistically significant level. Ekman *et al.* [177] found that physiological measures could distinguish some negative emotions such as fear from others like anger, for example in the case of anger , a higher skin temperature is observed than that for fear and sadness [61].

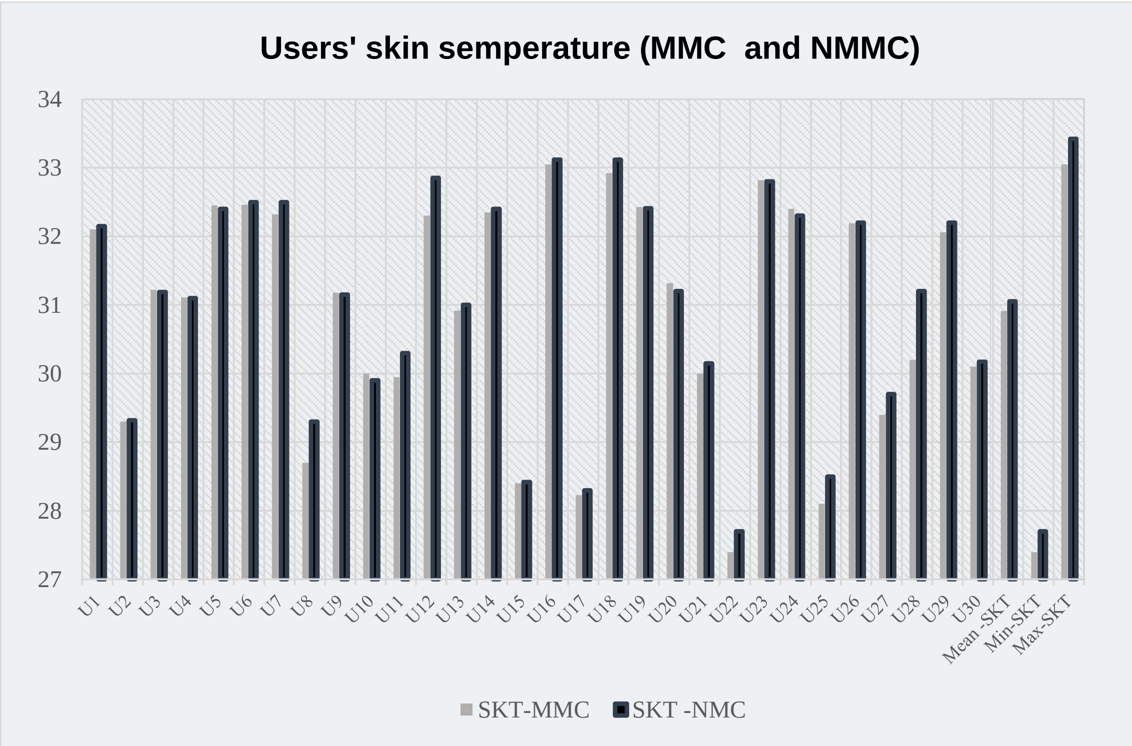


Figure 3-20: Users skin temperature (MMC and NMMC).

3.6.3.2 *SAM (Self-Assessments) Results*

Users' affective state towards the experimented conditions in terms of pleasure, arousal and dominance were evaluated using Self-Assessment Manikin (SAM) techniques. SAM is a non-verbal assessment tool, which is cross-cultural and easy to administer and understand.

Users' ratings were analysed and the test results are displayed in Figure 3-21. It was more negative and the overall rating was above neutral and slightly negative. A two related samples Wilxocon ranked test (Z=-4.858, P=.000) shows that the difference in the percived valence is significantly high, Figure 3-22 shows exactly the trend of users' emotions towards the the tested conditions. The layout of the MMC graph shows that users ranged from neutral to moderately positive with a great majority considering it to be moderately positive. The area covered in the graph indicates that agreement within users' ratings was quite high. For the NMMC, users alternated between negative and slightly negative emotions; the widely held views are placed between neutral and slightly negative with a tendency towards neutral. In their self-assesment users reported a positive emotion towards the MMC e-learning interface; their overall view is considered moderately positive; however, for the NMMC overall, the obtained results clearly show that the presence of multimodal communication metaphors was regarded as emotionally positive. With regards to arousal, the same test (-3.081, P=.002) indicated a significant statistical difference between users' ratings for both of the tested conditions. Users felt their emotions were stronger towards the MMC.

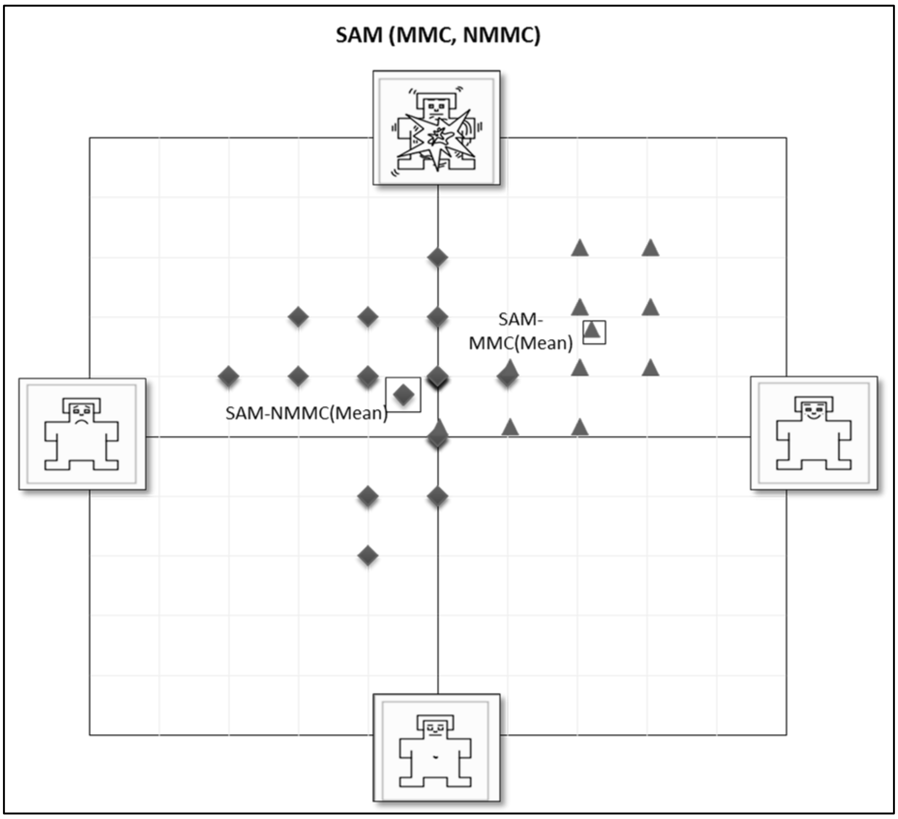


Figure 3-21: SAM result (MMC and NMMC).

Users' ratings were analysed and the test results are displayed in In their self-assesment users reported a positive emotion towards the MMC e-learning interface; their overall view is considered moderately positive; however, for the NMMC it was more negative and the overall rating was above neutral and slightly negative. A two related samples Wilxocon ranked test (Z=-4.858, P=.000) shows that the difference in the percived valence is significantly high. Figure 3-22 shows exactly the trend of users' emotions towards the the tested conditions.

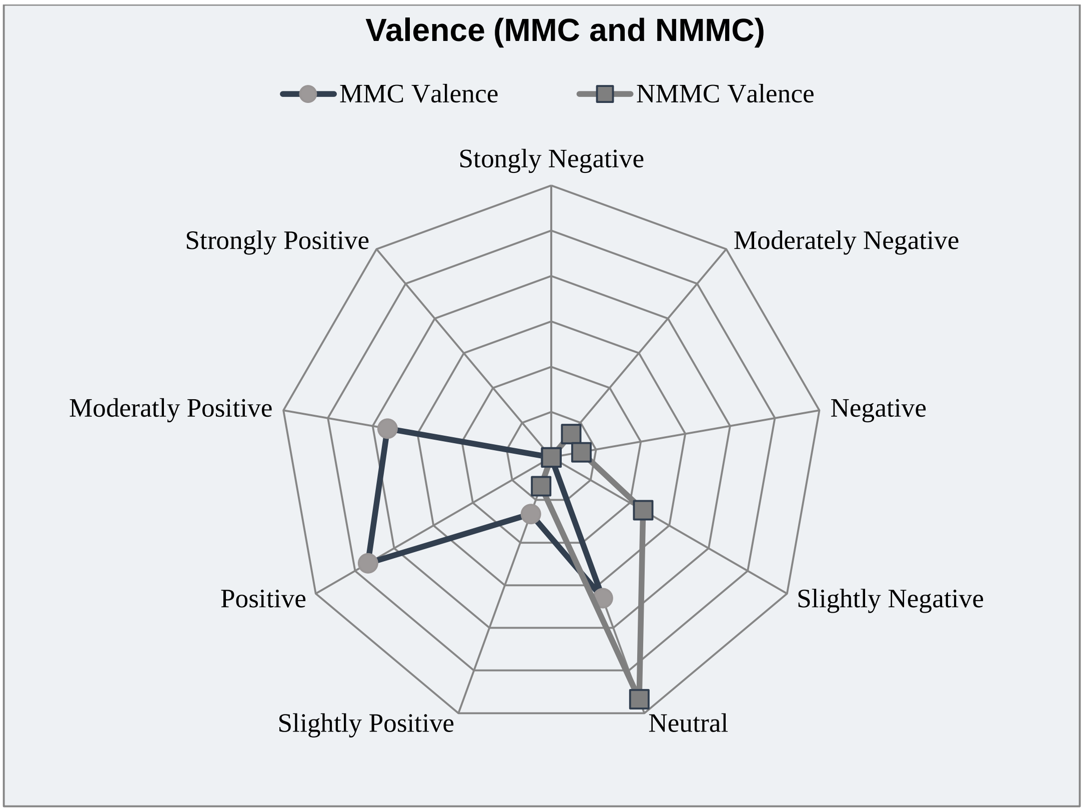


Figure 3-22: Emotions valence (MMC and NMMC).

Figure 3-23 indicates clearly how they rated their arousal within each experimented condition. Users' rating of the MMC ranged from high to slightly high, with the majority giving moderately high ratings. The layout of the line graph and the smaller area it covers show clearly that there is considerable agreement between users in their rating of the percived arousal. With regards to the NMMC, user rating fluctuated between low and moderately high with the majority rating it as slightly high. The shape of graph and area it covers indicate that users were moderately varied and there was some difference within subjects in their rating of the percived arousal.

Overall users were emotionally moved by the tested conditions although they were more aroused by the MMC; it is possible to conclude here that the noted increase was due to the inclusion of the multimodal communication metaphores in the MMC.

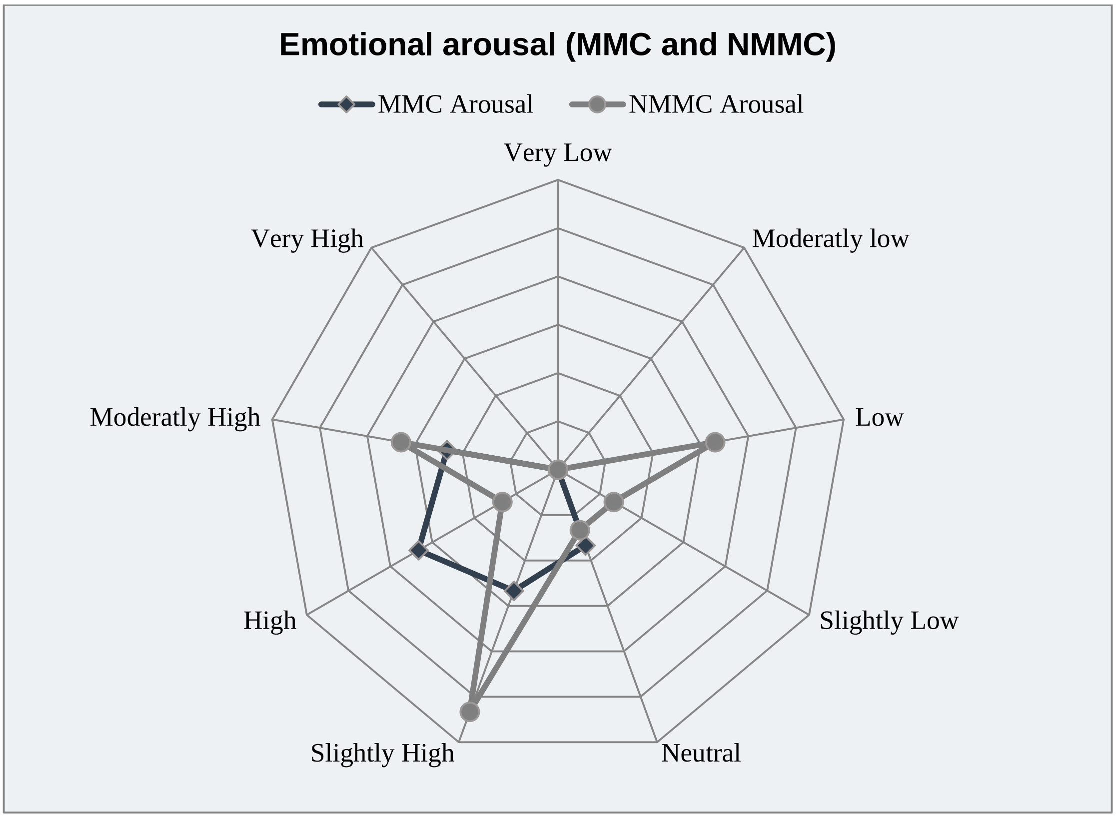


Figure 3-23: Users perceived arousal levels.

In terms of dominance, the results show that the majority of users felt they were in control of their emotion over the tested conditions(refer to Figure 3-24). The percentage for the MMC was 79.5% and 65.8% for the NMMC. A two related samples Wilcoxon ranked test result (Z=-3.081, P=.000) confirmed that the difference between users' rating within the dominance category reached a significant level. Figure 3-25 shows that dominance rating for the MMC ranged between users feeling they were in-control to feeling they were moderately in control. The direction of user rating is moving toward feeling in-control. Concerning the NMMC, users' rating of dominance by the condition ranged between neutral and in-control with a tendency towards neutral and slightly in-control directions. A high agreement between users in their rating of both conditions is indicated by the layout of the graphs and size of the covered area. Overall, users felt more in control while using the MMC when compared with the NMMC.

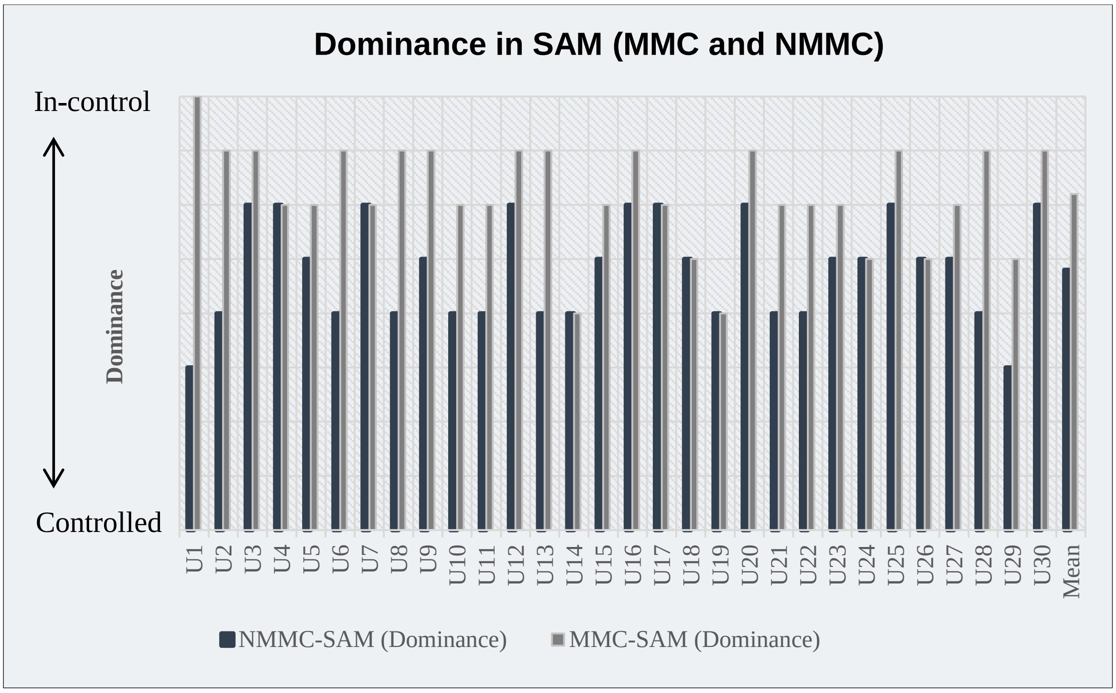


Figure 3-24: Dominance rating (MMC and NMMC).

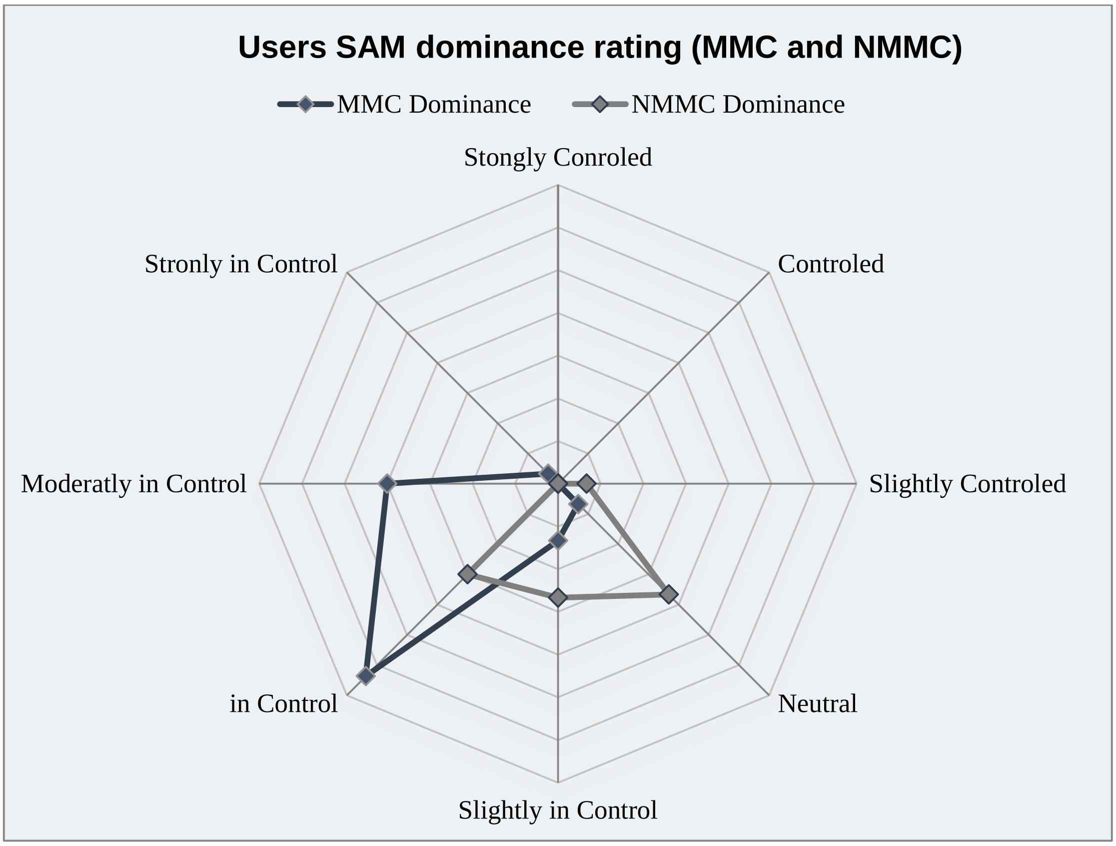


Figure 3-25: SAM dominance ratings (MMC and NMMC).

3.7 DISCUSSION

The current study empirically investigates the performance of e-learning user interfaces in the presence of multimodal communication metaphors (MMC) in comparison with the performance in their absence (NMMC). It is also concerned with factors, which affect learners' performance such as learning task type (recall and recognition) and complexity (easy, moderate and difficult). The study data was collected using subjective as well as objective techniques, employing questionnaires, observation, and biofeedback. The study results were analysed and compared using the within three areas:

- Usability evaluation: e-learning interfaces usability was evaluated in terms of user satisfaction, efficiency and user satisfaction.
- User experience evaluation: The experimental interfaces were evaluated in terms of their pragmatic quality, hedonic qualities and attractiveness.
- Emotion: Users' feelings and emotional reactions were evaluated while using the interface.

Accordingly, the results of the experiment will be discussed to reveal any contributions made by the multimodal metaphors in those three areas.

3.7.1 Usability Results

3.7.1.1 *Efficiency (task answering time)*

The first hypothesis states that the MMC would be more efficient than the NMMC in terms of time taken by the users to answer the presented tasks. Figure 3-5 (C) shows that using multimodal communication metaphors in the experimental condition (MMC) reduced the time spent by users on answering the presented tasks in comparison with the time they needed to accomplish the similar task using the control condition (NMMC). During the experiment, users' attention to the NMMC was split, as they had to switch eye contact regularly between texts and graphical illustrations to understand the presented topic. On the other hand, when using the MMC, they were able to maintain visual contact with the illustration and pay attention to the topic being presented. The presentation was given by a facially-expressive avatar, recorded speech narrations and auditory icons, alerting earcons and classical background music. Users seemed to enjoy the demonstration part where there was classical background music. The addition of multimodal communication metaphors helped the users to instantaneously utilize both visual and auditory channels which distributed the cognitive load [199] and reflected an improvement in task answering time therefore, H1 is fully accepted. The results are in favour of the hypothesis and therefore it is accepted. With regard to the second hypothesis, The results, displayed in Figure 3-5 (A and B) show that the increment in task answering time was in line with the escalation of task complexity in both conditions; however when comparing the increase in both conditions it is noted that it was less severe in the experimental condition at all levels of complexity. This shows that the use of multimodal communication metaphors is very useful when presenting complex information. As a task's complexity as well as cognition load increases more information is delivered and the available cognition resources are reduced. Multimodal communication metaphors could help increase the working memory by splitting the task processing load with the use of visual and auditory channels [199]. The experiment results indicate that the efficiency of the experimental condition increased with task complexity due to the inclusion of the multimodal communication metaphors; therefore, the hypothesis H2 is accepted.

3.7.1.2 *Effectiveness (task answering correctness)*

Concerning the third hypothesis, H3 states that the MMC will accomplish higher effectiveness than the NMMC in terms of correctly answered questions. From the results displayed Figure 3-8 (C), it is clear that the NMMC outperformed the NMMC in terms of correctly answered tasks and therefore H3 is accepted. Concerning the fourth hypothesis, H4 states that, the MMC would be more effective in performing tasks with higher complexity and cognition load. The results depicted in Figure 3-8 (A), clearly show that the MMC was steadily more effective than the NMMC on all complexity levels. For task type, the results in Figure 3-8 (B) show that the MMC maintained the same level of efficiency in answering recall and recognition tasks on all levels of complexity; therefore, H4 is fully accepted.

3.7.1.3 *User satisfaction*

Regarding the fifth hypothesis, H5, states that the MMC would yield a higher user satisfaction than the NMMC. The results displayed in Figure 3-9 representing SUS results, show that users were more satisfied with the MMC in comparison with the NMMC. Furthermore, the results of the additional (condition specific) satisfaction statements depicted in Figure 3-11 clearly show that the users' rating was in favour of the

MMC and that they were less satisfied with the NMMC; therefore, H5 is fully recognized and accepted.

3.7.2 User Experience

User experience was evaluated using AttrakDiff instrument [15] within three dimensions: experienced pragmatic and hedonic qualities and attractiveness. The results are discussed accordingly in the following subsections.

3.7.2.1 *Pragmatic quality*

The sixth hypothesis, H6 states that users would experience a higher pragmatic quality with the MMC. The results depicted in Figure 3-15, show that users' evaluations of the MMC in terms of its pragmatic quality was classified as "desired", and all users labelled it with high confidence as a pragmatic quality interface. For the NMMC, users' rating of its pragmatic quality was "neutral" with lower confidence, as users were not in high agreement in their rating; consequently, there is an urgent need for improvement and the NMMC was not classified as a product with a pragmatic quality. Users considered the MMC more manageable, predictable, clearer, more straightforward and more practical; however, they associated the NMMC with simplicity. It is noted that users rated the NMMC to be more technical as opposed to their rating of the MMC, which they considered to be more human and less technical. The result could be related largely to the inclusion of multimodal communication metaphors having humanized, to some extent, users' interaction with the tested condition (see Figures 3-16 and 3-17). Regarding overall user experience with the tested conditions and their ability to meet users' targeted goals in terms of the pragmatic quality, the MMC was more desired than the NMMC and was able to meet users' goals; therefore, H6 is accepted.

3.7.2.2 *Hedonic Quality*

Hedonic quality has two sides, user stimulation and identification with presented conditions. The results are depicted in, Figures 3-16 as well as 3-17 and discussed below.

Hedonic Quality-Identification: Users' evaluation of HQ-I placed the MMC in the above average region and they identified well with it; the MMC encouraged users, inspired their enthusiasm and awakened their desire to learn. Consequently, users identified themselves well with the MMC condition, which they have rated it to be optimal. With regard to the NMMC, users placed the condition in the average region. It was just capable of meeting ordinary levels so in order to win users and make them continue to use such a system improvements are necessary. The results clearly confirm the correctness of H7, which stated that users would experience higher identification with MMC. As a result, the hypothesis is accepted.

Hedonic Quality-Stimulation: With respect to hedonic quality stimulation, users' ratings positioned the MMC in the above average region. Users were stimulated by it, their curiosity was awakened and they were motivated and inspired. Largely, in terms of stimulation users classified it as optimal. However, their ratings of the same category in the NMMC positioned it lower than the average region, as the NMMC did not have a stimulating effect on them. The lack of stimulation resulted in lack of motivation to use the system; it is worth mentioning that, should a system of similar pragmatic quality be available users would happily migrate towards it, so the NMMC condition is in serious need of amendment. Overall, users were more stimulated by the MMC; therefore, H8, which predicted this finding, is proven true and is fully accepted. Furthermore as the results showed, the MMC provided a better hedonic user experience than the NMMC, therefore the H9 is accordingly accepted.

3.7.2.3 *Attractiveness*

According to users' ratings in Figures 3-16 and 3-17, The MMC attractiveness value was placed above the average region. The total impression of the condition is highly attractive. With respect to users' rankings of the same category in the NMMC the attractiveness value was situated lower than the average region; the general impression of the conditions is unattractive. Users thought the MMC was more appealing, inviting, likable, pleasant, and attractive; this finding confirms what was stated in H10, therefore it is considered proven and accepted. Overall, the inclusion of the multimodal communication metaphors in the MMC has improved the overall user experience; consequently, H11 is accepted.

3.7.3 User Affective state

3.7.3.1 *Biofeedback*

Skin Conductance: Users' electro-dermal activities were monitored by recording changes in skin conductance; the results (see Figures 3-18 and 3-19) indicated that users' skin conductance while using the MMC was higher compared to the NMMC.

The statistical difference in users' skin conductance in both conditions reached a significant level. Stronger stimulation produces stronger change in skin conductance. The general increase in electro-dermal activity is reflected as an increase in skin conductance. The observed increase can be associated with increase in arousal and attentiveness [134]. The results clearly show that users were more stimulated by the MMC. They encountered a higher number of stimuli, the interpretation of changes in electro-dermal activities should be made in relation to other measurements to decide, whether the perceived increase was positive or negative. The results are clearly in favour of H15 and therefore, it is accepted.

Skin Temperature: Changes recorded in users' skin temperature while they were using the experimental conditions are displayed in Figure 3-20. The overall result shows that there was a slight increase in skin temperature in the NMMC in comparison with the MMC. A statistical test showed that the difference reached a significant level. Increase in skin temperature can happen due to experiencing either positive or negative emotional episodes ,even though such increase in the case of negative emotions could distinguish sadness or fear from anger [200]. Accordingly, with reference to the subjective assessment, the observed increase in users skin temperature while using the NMMC was an indicator of experiencing negative emotions. Utilizing SAM assessments, users' affective state was evaluated within three dimensions including valence, arousal and dominance; the results are presented here and discussed accordingly.

3.7.3.2 *SAM (Self-Assessment)*

Valence: The term valence refers to the positive or negative emotion in reaction to a presented stimulus [61]. In their valence evaluation (see Figure 3-21), users regarded the MMC (experimental e-learning interface) to be more positive and in their overall view rated the condition as moderately positive. Conversely, their overview of the NMMC (controled e-learning interface) was more negative and generally their rating ranged between neutral and slightly negative. The statistical tests conducted confirmed that the difference between their ratings for both conditions was significantly high. Users' ratings were analysed and the test results are displayed in Figure 3-21.

In their self-assesment users reported a positive emotion towards the MMC e-learning interface; their overall view is considered moderately positive; however, for the NMMC it was more negative and the overall rating was above neutral and slightly negative. A two related samples Wilxocon ranked test (Z=-4.858, P=.000) shows that the difference in the percived valence is significantly high. Figure 3-22 accurately shows the trend of how the valence was percived. The layout of MMC graph line demonstrates that the majority of users rated the condition as moderately positive. On the other hand, users' ratings for the NMMC fluctuated between negative and slightly negative emotions; the NMMC graph line layout and the area it covers indicate that users' views ranged between neutral and slightly negative with a tendency towards neutral. The results are clearly in favour of the MMC, therefore H11, which states that the MMC will induce more positive valence, is accepted, and it is safe to conclude that the inclusion of multimodal communication metaphors (in the experimental condition) resulted in a more positive valence.

Arousal: The term arousal in physiology and psychology refers to the state of being active when presented with a stimulus ,arousal indicates the strength of particular emotion [106] and results in an increased skin conductivity [107]. Arousal might be positively or negatively valenced according to the way in which they are perceived. Users' subjective evaluation of the tested conditions was meaured using SAM scale and the results are presesented in Figure 3-23 show how users rated their arousal, which was stronger towards the MMC than the NMMC; thus, the difference in the rating of arousal between the two conditions is statistically significant. Figure 3-26 clearly shows how the rating of arousal was different, with that of the MMC extending from high to slightly high with a major tendency with significant agreement toward moderately high. User rating for the NMMC varied and fluctuated between low to moderately high with a tendency toward slightly high. Overall, users were more aroused by the MMC; therefore, H13, which states that users will be more alert and engaged while using the MMC, is partially accepted and it is possible to conclude that users' perceived arousal increases were due to the inclusion of the multimodal interaction metaphores.

Dominance: The term dominance can be viewed as the extent to which users assume control over reaction to move backward (are demotivated) or forward (are motivated) when presented with stimuli ,it can be described as a dominance-submission for a person in a given situation [109]. The results shown in Figure 3-24 indicate that the majority of users felt they dominated the tested conditions. The percentage was 79.5% in the MMC, followed by 65.8% in the NMMC. The statistcal difference was significant between users' ratings for both conditions. The results show that dominance rating for the MMC ranged between being in control to being moderately in control with some tendency toward assumed being fully in control. In the NMMC, users' rating of dominance by the condition ranged from neutral to slightly assumed control. Overall, users felt more in control and engaged while using the MMC compared to the NMMC. With reference to the dominance and arousal testing results it is hence viable to conclude that H14 is fully accepted. The results of testing the aspects of users' affective state (valence, arousal and dominance) clearly show that users' affective state was more positive toward the MMC compared to the NMMC. Therefore, it is concluded here that the inclusion of the multimodal interaction metaphors made a positive impact upon user affective state. Consequently, H16 is fully accepted.

3.7.3.3 ***Holistic overview***

Assessing users' affective state while they were interacting with the experimental conditions required a holistic overview taking into account the following variables:

- User satisfaction (usability).
- User experience in terms of attractiveness as well as pragmatic and hedonic qualities.
- User biofeedback.
- Arousal, valence and dominance (SAM).

The mentioned variables, depicted in Figure 3-26, show that user ratings were in favour of the MMC.

Users were more positive in rating their satisfaction with the MMC; their overall experience was very positive and they felt it was highly attractive. With regards to biofeedback, skin conductance was higher indicating a higher level of stimulation; skin temperature was slightly lower which indicates that the excitement was positive as slightly negative emotions are accompanied by increase in skin temperature [201]. Concerning SAM scale the results correlated well with the other variables; user rating of valence-arousal indicates that they were positively aroused. Concerning dominance, users were in favour of the MMC and they felt more in control; they felt they would like to continue using it.

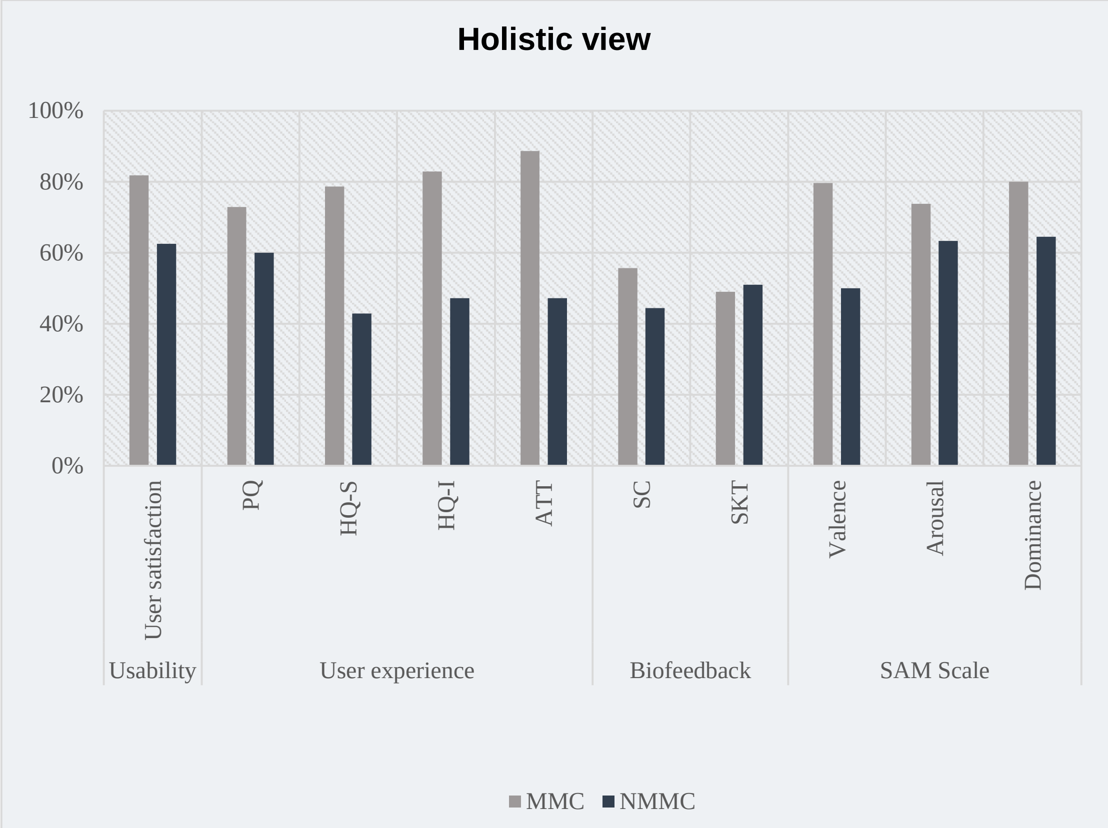


Figure 3-26: The overall users' affective state.

3.8 SUMMARY

The first experimental study aimed at investigating the impact of including multimodal communication metaphors in e-learning interfaces on the perceived system usability, user experience and emotions. Based on the measured variables a total number of 16 hypotheses were formulated accordingly to reveal the influence of using multimodal communication metaphors on e-learning interfaces. To meet the objectives of the study and test the hypotheses, an experimental e-learning platform was designed and implemented. The experimental platform included two conditions and adopted a within-subject design whereby the same user experimented with both conditions; the assessment criteria involved measuring conditions' usability and users experience and emotions. The data obtained from both tested conditions were analysed and compared. The results undoubtedly confirmed that, including multimodal communication metaphors vividly improved interface usability, enhanced user experience, and induced positive emotions. The experimental work results are summarized in the triple evaluation traffic light (see Figure 3-27).

Evaluation Category	Aspects	MMC	NMMC
Usability	Efficiency	✔	!
	Effectiveness	✔	!
	Satisfaction	✔	✖
	Learning Performance	✔	!
User Experience	PQ	✔	!
	HQ-I	✔	✖
	HQ-S	✔	✖
	ATT	✔	✖
User Affective State	SC	✔	✖
	SKT	✔	✖
	Valence	✔	!
	Arousal	✔	!
	Dominance	✔	✖

Figure 3-27: The triple evaluation traffic light.

However, further investigation is needed to uncover the role each communication metaphor played in the perceived improvements and to answer the following questions:

Which communication metaphor contributed more to the results?

How can they be wisely utilized?

How often should they be employed?

CHAPTER 4

4 EXPERIMENTAL PHASE II: MULTIMODAL COMMUNICATION METAPHORS DIFFERENTIATION

4.1 INTRODUCTION

With reference to the perceived enhancement in the first experimental phase, this experiment aims to distinguish the impact of each communication metaphor used on the interface usability as well as user experience and affective state. It looks at which metaphor had a more positive impact on users and how they could be integrated to achieve optimal results. Accordingly, an e-learning experimental platform included five experimental conditions (auditory icons, earcons, music, facially expressive speaking avatar and recorded speech) in addition to a control condition (NMMC) was constructed. This setup was used to serve as a basis for the investigation. To insure the validity of the results and make sure that obtained results were not due to any prior experience, the second experimental phase involved a new group of 33 users (only 30 of them successfully completed the experiment and the rest were excluded from the evaluation). The group of users presented an opportunistic sample from De Montfort University, they were randomly selected volunteers and none of them had participated in the first experimental phase. Conditions' usability, user experience and affective state were evaluated using subjective and objective approaches.

Due to the size of the experiment and the purpose of clarity, we have chosen to present this experimental phase in three chapters:

- Chapter 4 presents the conducted experiment in addition to the aims, objectives, hypotheses and results of the system usability evaluation.
- Chapter 5 presents the user experience evaluation aims, objectives, hypotheses and results.
- Chapter 6 introduces the evaluation aims, objectives, hypothesises and the user affective state evaluation results.

4.2 AIMS AND OBJECTIVES

The principal aims of the experiment are to:

- Examine the influence each communication metaphor has on system usability (see Section 6.4 for the aims and objectives in details).
- Investigate the impact of each communication metaphor on the overall user experience (see Section 5.1 for the aims and objectives in details).
- Examine the impact of each communication metaphor on user affective state (see Section 6.1 for the aims and objectives in details).

Satisfying the aims of this experiment requires meeting the following objectives:

- Formulate experimental hypotheses (see Sections 4.3, 5.2 and 6.2).
- Create an e-learning experimental platform with six conditions to execute the empirical investigation using the following setup:
    - Auditory experimental condition, the condition consists of auditory icons with text and graphics (see Section 4.5.1).
    - Avatar experimental condition, the condition contains facially-expressive speaking avatar with text and graphics (see Section 4.5.2).
    - Earcons experimental condition, the condition consists of earcons combined with text and graphics (see Section 4.5.3).
    - Music experimental condition, the condition includes classic music background with text and graphics (see Section 4.5.4).
    - Speech experimental condition, the condition includes recorded speech narrations in combination with text and graphics (see Section 4.5.5).
    - Control condition, the condition consists of text and graphics only (see Section 4.5.6).
- Evaluate conditions usability (see Section 4.6).
- Evaluate the user experience (see Section 5.4).
- Evaluate the user affective state (see Section 6.4).

4.3 HYPOTHESES

It is anticipated that the utilised communication metaphors will influence system usability, user experience, learning performance, affective state and a series of hypotheses have been formulated accordingly:

- Usability hypotheses (see Section 4.6.2).
- User experience hypotheses (see Section 5.2).

- User affective state hypotheses (see Section 6.2).

4.4 LEARNING TOPICS AND CONTENTS

Selecting the learning topic took into account the fact that the targeted users are university students majoring in Computing, Engineering, Law, Health, and Social Sciences. To ensure data validity, a non-familiar learning subject was chosen (Earth Science) in order to give all participants an equal chance and provide consistent and accurate results. All topics presented were related and had the same level of length and complexity. The topics were:

- Volcanoes (auditory condition).
- The hydrosphere (music conditions).
- The Earth's internal structure (avatar condition).
- Rock types (earcons condition).
- The rock cycle (speech condition).
- Mining geology (NMMC).

4.5 EXPERIMENTAL CONDITIONS

4.5.1 Auditory Icons Experimental Condition

Encompasses a basic graphical user interface with animation accompanied by auditory icons. Auditory icons are the main communication metaphor being investigated in this part of the experiment. Auditory icons are naturally-occurring sounds extracted from a video recording of a volcanic eruption. The communication metaphor is played in the background while users navigate through the lesson (see Figure 4-1 A). The auditory icons are expected to make the lesson realistic and help users feel the real life situation.

4.5.2 Avatar Experimental Condition

Consists of a graphical user interface combined with a facially expressive avatar (see Figure 4-1 B). Avatar is the main communication metaphor being utilized to deliver the learning topic in this experimental condition. It is anticipated that it will have a positive influence on users as it replicates a real-life classroom experience.

4.5.3 Earcons Experimental Condition

This condition presents the learning material using text and graphics and employs earcons with text boxes to highlight the main ideas and keywords of the lesson.

Screen captures from the tested conditions

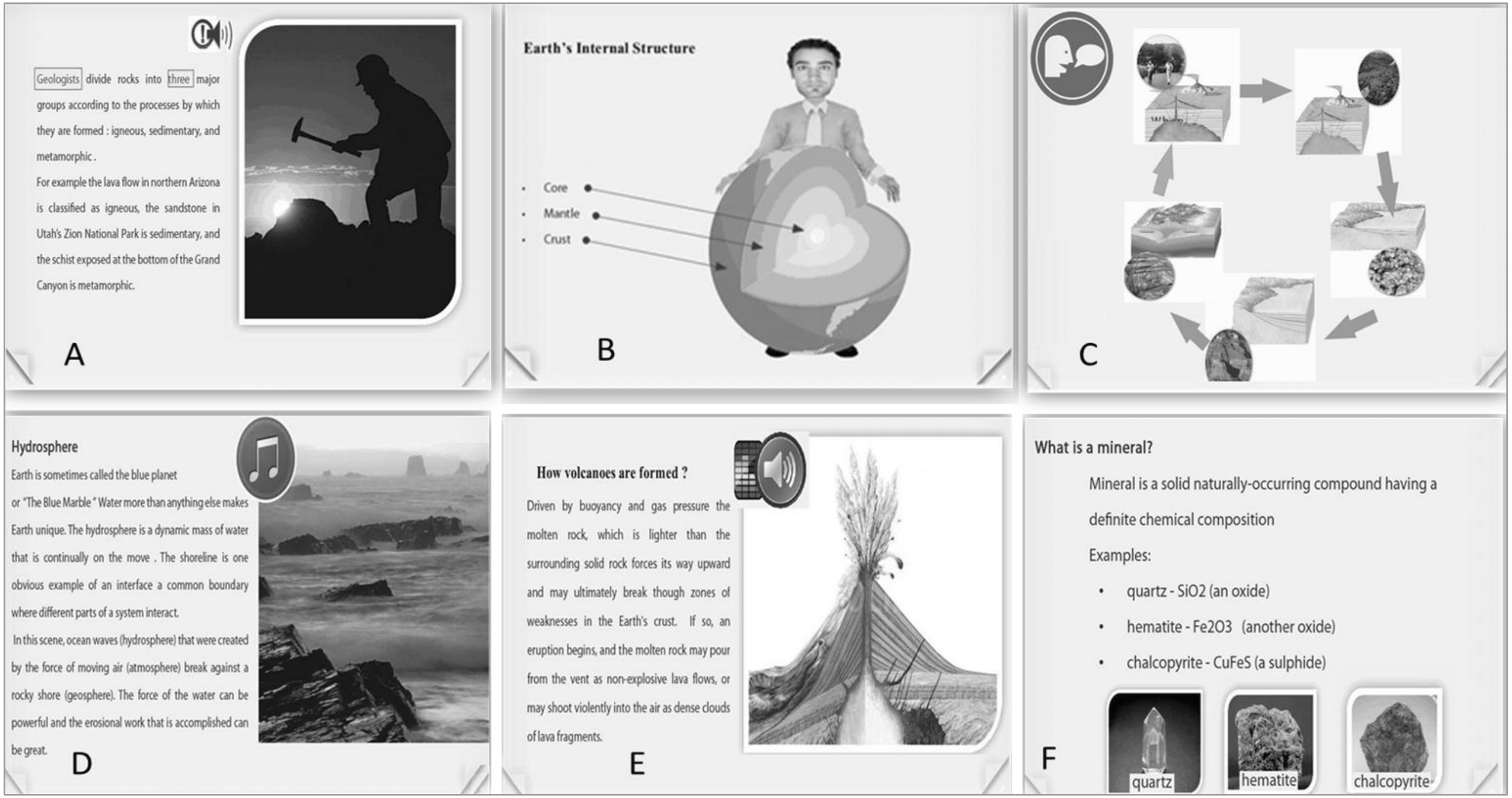


Figure 4-1: Screen captures from the tested conditions.

Users are requested to hover the mouse pointer over the written text and as they do so the keywords appears accompanied by an earcons alert and highlighted text boxes. Earcons are the main communication metaphor being investigated; they are expected to help users identify the main ideas presented and the essential keywords of the learning topic (see Figure 4-1 C).

4.5.4 Classic Music Experimental Condition

This condition was constructed with a combination of a graphical user interface and classic music. Music is the interaction metaphor being investigated in this section. The interaction metaphor will be used as a background sound during the course of interaction between the users and the lesson. It is assumed that users will benefit from this metaphor, as it will enable them to relax and enjoy the topic (see Figure 4-1 D). The music used was composed in 1725 by Antonio Vivaldi and named La Primavera (Spring), it is part of his most famous work, the Four Seasons [202 and 203]. La Primavera was used by researchers to study the relation between music and emotion [204 and 205].

4.5.5 Speech Experimental Condition

In this condition, recorded speech is the main communication metaphor used to deliver the learning topic. It is combined with a synchronised graphical illustration and presented in a timely manner to achieve the best possible presentation (see Figure 4-1 C). It is presumed that this approach will help the users concentrate on the voice of the presenter as well as the illustration without being interrupted by having to read any text.

4.5.6 NMMC (Control) Condition

This condition employs a typical graphical user interface to deliver the learning topic. It consists of text and graphics for illustration purposes (see Figure 4-1 E). This condition is controlled and to be used as a benchmark for comparing experimental conditions.

It is presumed that with this approach the multimodal platform will outperform the other experimental conditions and positively influence the condition's usability, users' experience and affective state.

4.5.7 Tasks and Variables

In this experiment, there are a number of controlled and non-controlled variables, as listed in Table 4-1.

The study variables		
Type	**Variables**	**Attributes**
Controlled variables	Procedure	All the users were treated equally and to insure that: • The same procedure, hardware and software were used for the entire experiment with all the users. • Biofeedback device timing was synchronized with the clock of the machine running the experimental platform.
	Learning topics	All the learning topics presented in the experiment had the same: • Theme, length, complexity level and presented topic.
	Platform	• Communication metaphors used to deliver the learning contents of each experimental condition were the same for all the users. • The experimented conditions were randomly presented and without any preference.
	Tasks	• Number of tasks and complexity levels were the same for all users and tested conditions. • Number of allowed trials (all users were limited to one round to perform presented task correctly). • All participants were first time users.
	Self-Report (Subjective measurements)	• Number of questions. • Type and objectives of questions. • Questions' length.
	Biofeedback (Objective measurements)	• The same sensor was used throughout the trial. • The same types of physiological signals were measured for all the users and in all of the tested conditions.
Dependent Variables	Time Spent	• Lesson completion time. • Task execution time (in terms of type and complexity).
	Task correctness	• Measured in relation to task type and complexity level.
	Usability	• Efficiency, effectiveness as well as user satisfaction and learning performance.
	User experience	• Hedonic as well as pragmatic quality and attractiveness.
	Biofeedback	• Skin conductance and temperature.
Independent Variables	Platforms and used conditions	Six condition were used in the experimental platform: • Auditory icons experimental condition. • Music experimental condition. • Earcons experimental condition. • Avatar experimental condition. • Speech experimental condition. • NMMC control condition.

Table 4-1: Experiment Variables

4.5.7.1 *Questionnaires*

The users were requested to complete a set of pre-session and post-session questionnaires.

4.5.7.2 *Users*

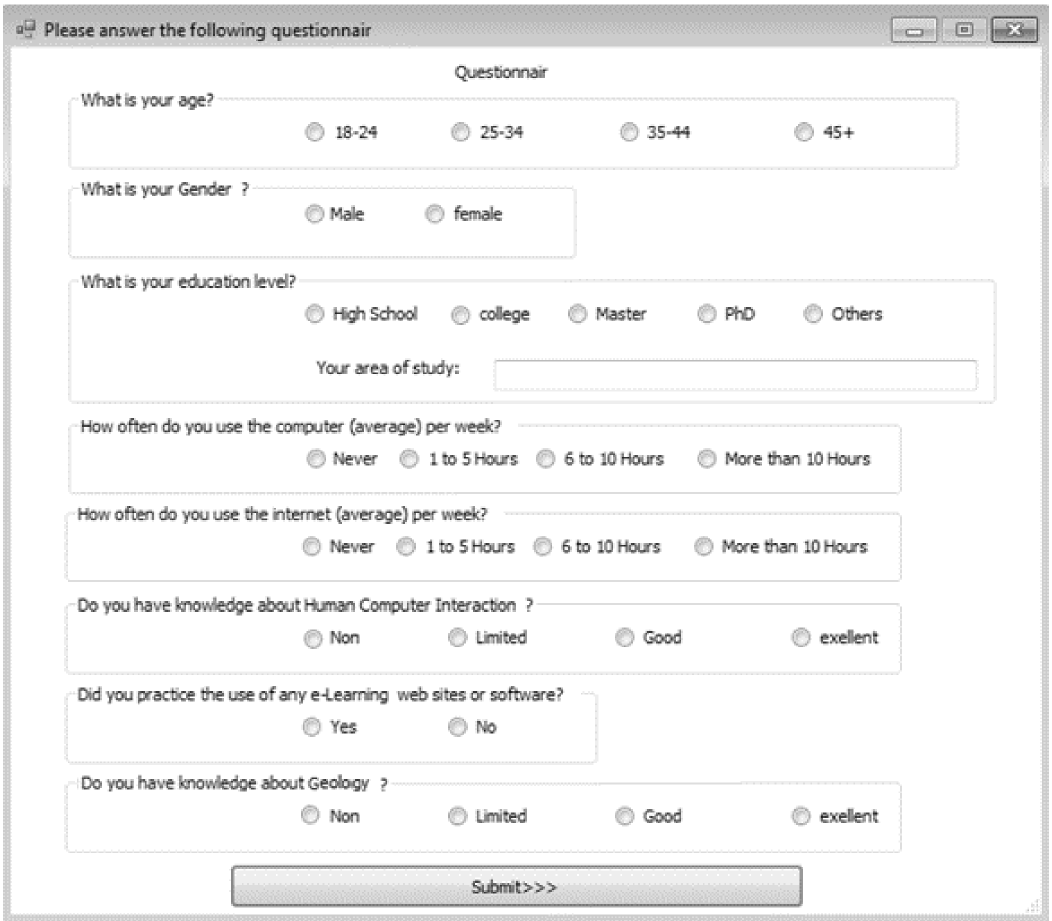
A total number of 33 users took part in the experiment. Two potential participants withdrew from the trial because of time constraints and other commitments; while another participant's physiological data recording was not complete, therefore their input was eliminated. That brought the total number of users who successfully completed the experiment to 30.

4.5.8 **Experiment Procedure**

All users were welcomed by the researcher and briefed about the experiment as well as their rights, privacy, data protection and university ethical code. The researcher made it clear to them that they poses the right to pull out from the trial at any point of time, without any obligations , the following steps were followed for each participant:

1. The researcher briefed the users about navigation of the platforms as well as the steps involved to complete the experiment.
2. Following that, all users were asked to wear the biofeedback device on the palm of their left hands for ten minutes prior to conducting the trial; this would establish a baseline measurement for each user.
3. Next, prior to starting the experiment, the researcher started recording all the desktop activities; the records would be used for data collection.
4. Users were requested to complete a questionnaire about themselves and their background; the information included their age, gender, education level, area of study, average use of computers and the internet, knowledge of the presented subject and average use of e-learning (see Figure 4-2).
5. Upon completing the questionnaire and submitting the answers, users were requested to press the control button of the biofeedback device to mark their exact starting time.
6. After marking the starting point, the user pressed next and proceeded to the first experimental condition which was randomly picked by the platform; this was done to ensure that there were no biased results and that all platforms had an equal effect

on the users (see Figure 4-3, which explains the navigation process of the experiment).


Please answer the following questionnair

Questionnair

What is your age?
18-24 25-34 35-44 45+

What is your Gender ?
Male female

What is your education level?
High School college Master PhD Others
Your area of study:

How often do you use the computer (average) per week?
Never 1 to 5 Hours 6 to 10 Hours More than 10 Hours

How often do you use the internet (average) per week?
Never 1 to 5 Hours 6 to 10 Hours More than 10 Hours

Do you have knowledge about Human Computer Interaction ?
Non Limited Good exellent

Did you practice the use of any e-Learning web sites or software?
Yes No

Do you have knowledge about Geology ?
Non Limited Good exellent

Submit>>>

Figure 4-2: Questionnaire to obtain users' information.

7. Once the user had completed the lesson, an instruction appeared on the screen asking them to press the biofeedback device's control button to mark the completion point, and then they continued to perform the recognition and recall tasks.
8. In the recall and recognition tasks, users were requested to answer two questions (easy and moderate) for each category in relation to the presented topic; once they had completed the fully performed the tasks immediate feedback was given.
9. Next, users proceeded to the usability feedback where they had to answer a set of five satisfaction questions related to the tested condition.
10. Then, users went on to use the AttrakDiff scale to reflect on their experience.
11. Then users were asked to assess their affective state using SAM.

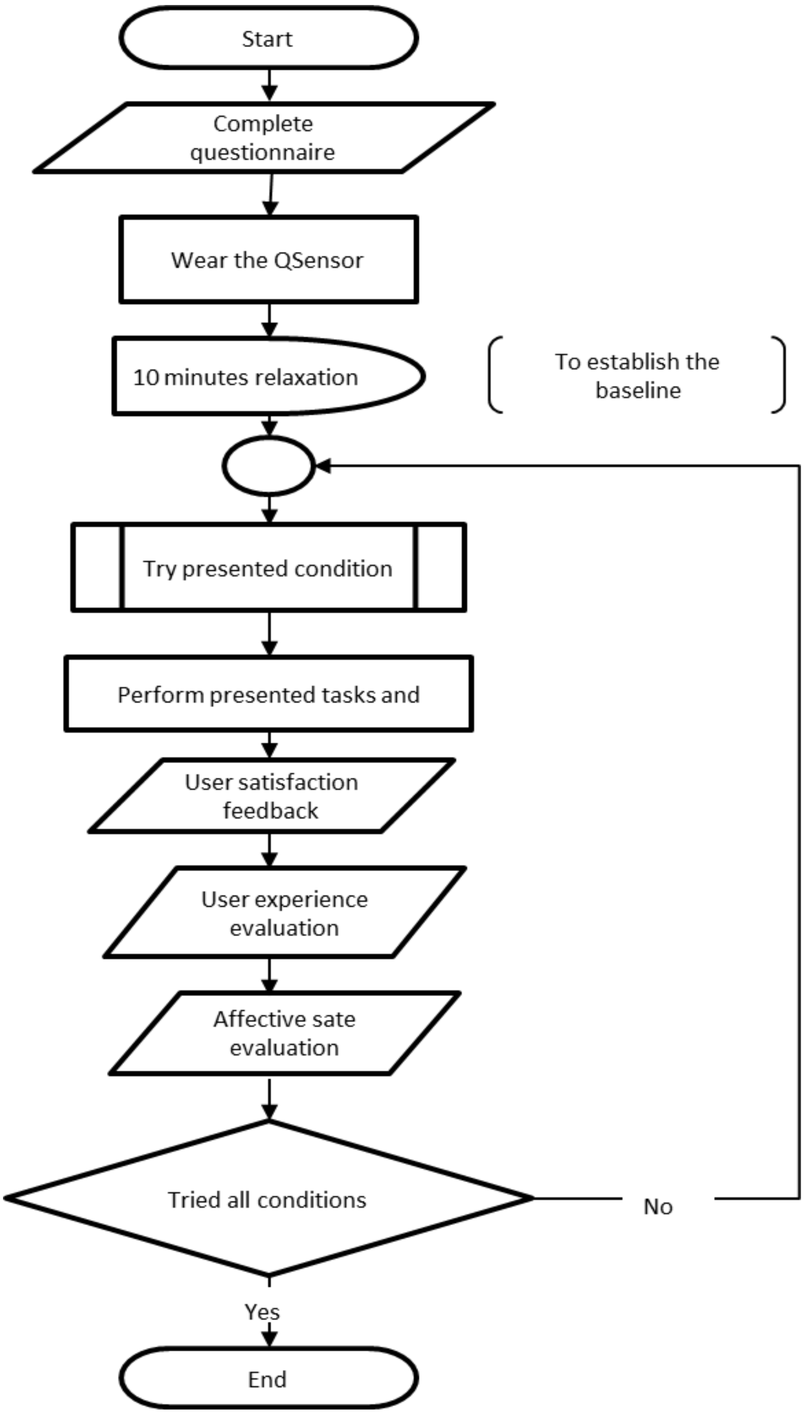


Figure 4-3: Trial navigation.

4.5.9 **Pilot Test**

In order to ensure the validity of the designed platform and identify any problems, which might arise during the actual trial, a pilot test was conducted with four volunteers. They tested the experimental platform on an individual basis and highlighted the following issues:

- Completing the experiment required a very long time.
- There were too many questions to answer for each tested condition.

Consequently, a revised and shortened version of the platform was built. The post-experiment satisfaction questionnaires were reduced for each condition, without compromising on their quality or objective.

4.5.10 Users' Profiles

All of the users were university students, 43.33% were undergraduates, 40% Master's students and 16.67% PhD students. Of these, 80% were males and 20% females; 43.33% were aged between 18 to 24 years, 30% were aged 25-34 years, 23.33% were 35-44 and3.33% were aged over 45 years (see Figure 4-4). All participants were regular computer users, the results indicated that 66.67% of them used computers for more than ten hours per week, 26.67% used them for between 6-10 hours a week and 6.67% used computers for less than 6 hours weekly.

Similarly, regarding the internet, more than 70% used the internet for more than ten hours per week, while 20% used it for 6-10 hours and 6.67% used it for less than 6 hours weekly. With reference to the research area and the presented topic, more than three quarters had no knowledge of presented topic, 16.67% had only limited knowledge and less than 7% had good knowledge in the area.

For the human computer interaction, half of the users had limited knowledge in the area, one third of the users had no prior knowledge, 13.3% stated that they had good knowledge and less than 7% described their knowledge as excellent.

4.5.11 Experimental Sessions

The time spent on the experiment can be divided into three sections, pre-experiment, actual experiment and post-experiment time. The experimental sessions lasted between 46:51 minutes and 67:20 minutes with a mean value of 56:00 minutes (see Figure 4-5, for the total accumulated experimental time).

On average, users spent 08:07 minutes on the auditory icons condition, 07:51 minutes on the avatar platform, 07:35 minutes on the earcons condition, 07:24 minutes on the MMC, 08:40 minutes on the music condition, 08:27 minutes on the NMMC and 07:52 minutes on the speech condition (see Figure 4-6).

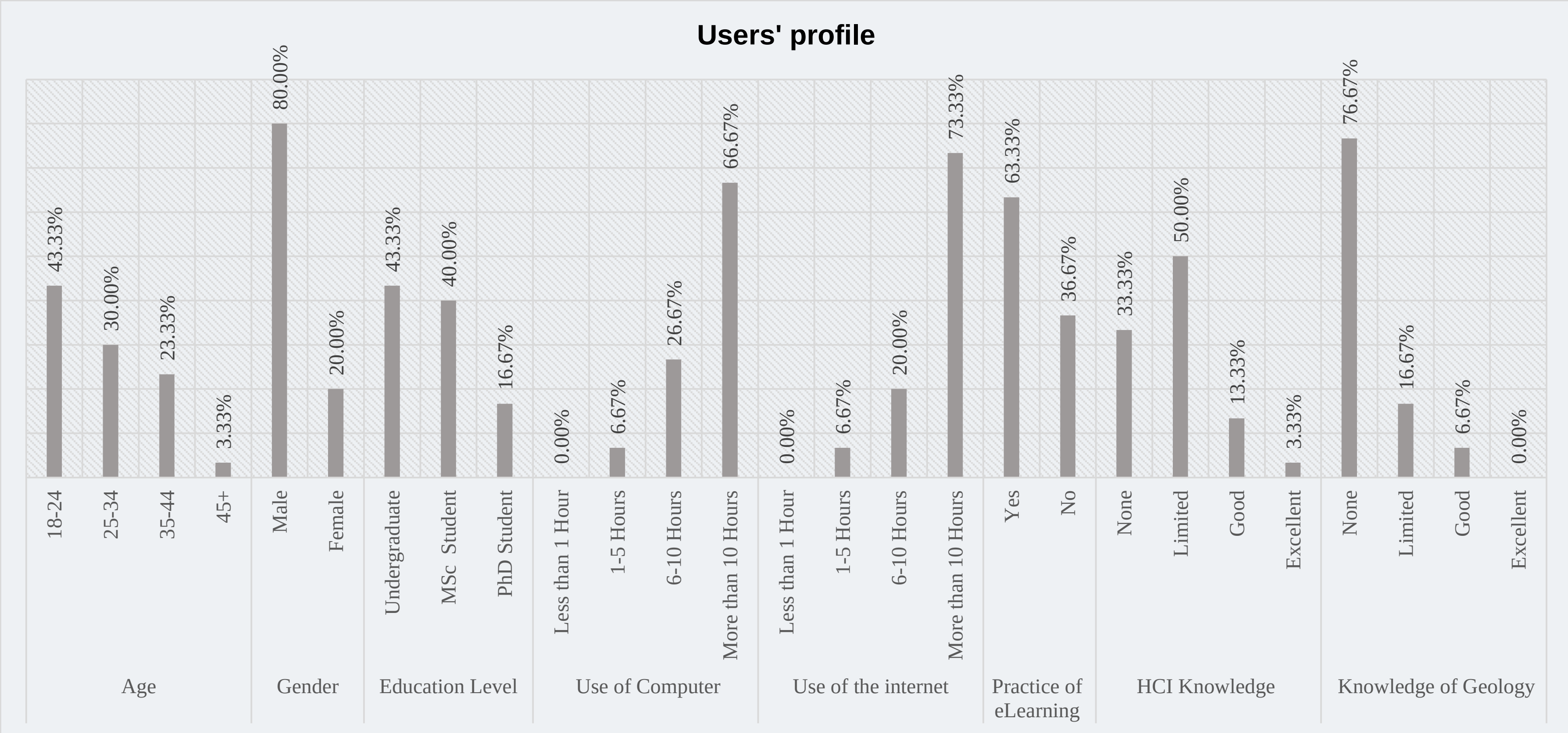


Figure 4-4: Users’ profile.

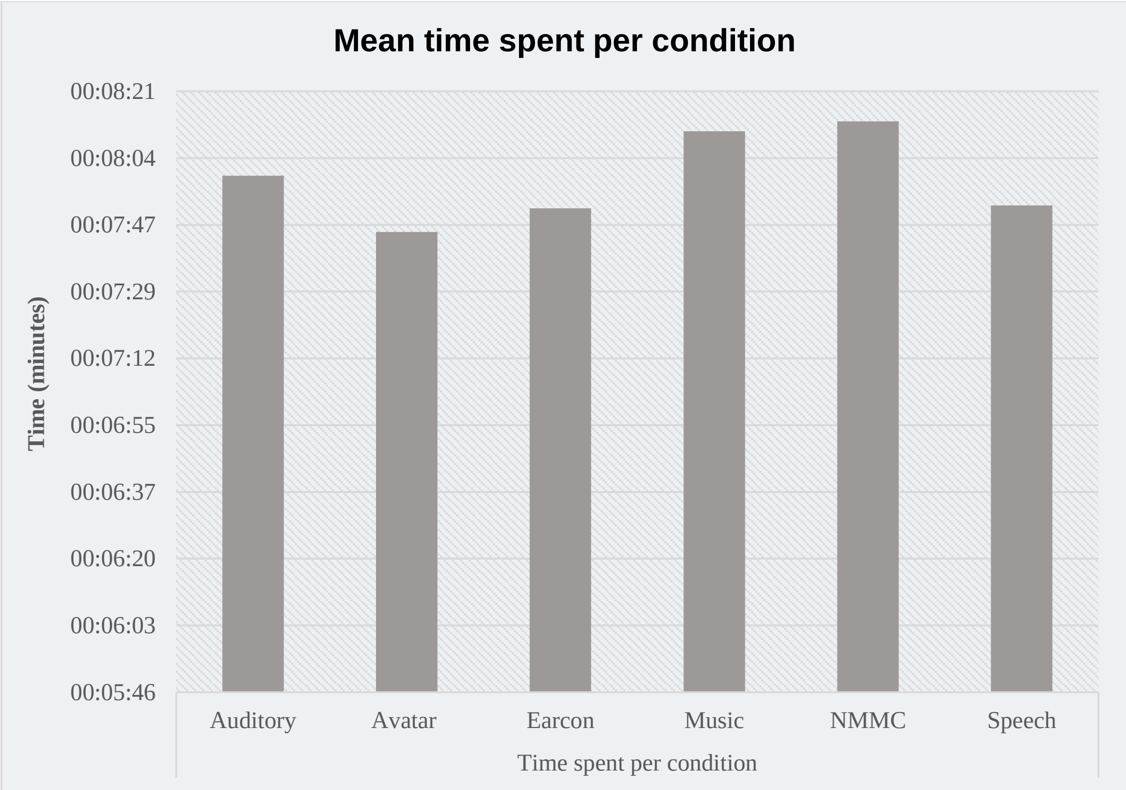


Figure 4-5: Mean time spent per condition.

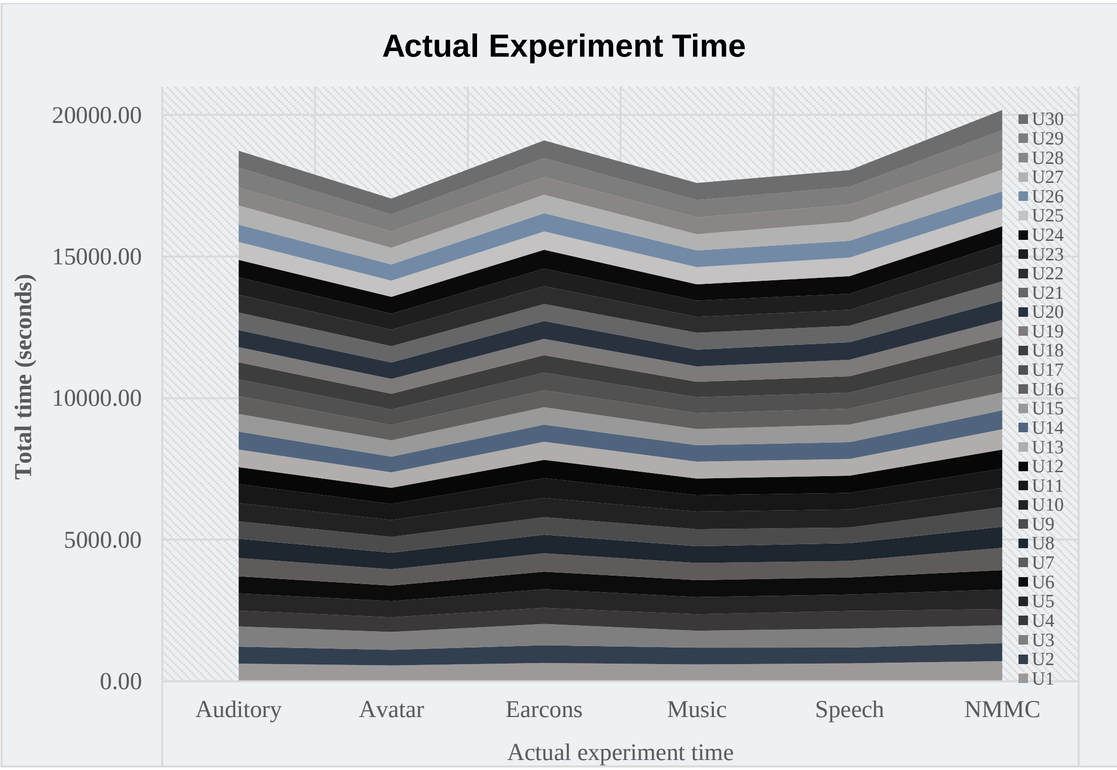


Figure 4-6: Accumulated time (spent by users for experimental sessions).

4.6 USABILITY EVALUATION (EXPERIMENTAL PHASE II)

4.6.1 Aims and Objectives

The principal aims of the experimental study concerning system usability are to examine the influence each communication metaphor has on system usability and accomplish the following:

- Try six experimental conditions where communication metaphors were individually employed.
- Examine the influence each communication metaphor has on system usability and find any significant differences between usability of the tested communication metaphors within the following areas:
    - Effectiveness.
    - Efficiency.
    - User satisfaction.
    - Learning performance.
- Identify which condition achieved the highest performance in terms of system usability.

To accomplish the aims of this experiment in investigating system usability, the following objectives have to be met:

- Formulate usability experimental hypotheses.
- Evaluate conditions usability by:
    - Measuring conditions' efficiency by measuring the time taken by users to accomplish the presented tasks.
    - Measuring effectiveness by calculating the percentage of correctly completed tasks.
    - Measuring users' satisfaction (using SUS combined with additional condition-tailored satisfaction statements).
    - Evaluating users' learning performance within each tested condition (by comparing learning time and outcomes).
- Statistically evaluate the results obtained from the experimental study to find out any significant difference between the tested condition and try to identify the condition achieved the highest usability score.

4.6.2 Hypotheses

It is anticipated that the utilized communication metaphors will influence the system usability, learning performance and users' satisfaction. Accordingly, the following hypotheses have been formulated based on the selected evaluation criteria and the measured variables:

H1: All the included communication metaphors will lead to improvements in the system efficiency.

H2: The experimental conditions are equally efficient in terms of the time required to execute the presented tasks.

H3: The inclusion of the various communication metaphors will improve the system effectiveness.

H4: The experimental conditions are equally effective in terms of the number of correctly performed tasks.

H5: The experimental conditions are equally effective and efficient in performing both recall and recognition tasks.

H6: Individually utilized communication metaphors are equal competent in terms of their impact on users' learning performance.

H7: The experimental conditions will receive equal preferences in terms of user satisfaction.

4.6.3 Usability Evaluation Results

4.6.3.1 *Efficiency*

Efficiency is measured by calculating the resources taken to achieve a particular goal. Participants in this experiment were requested to perform twelve tasks, six recall tasks and six recognition tasks. Furthermore, the tasks also varied in their complexity from easy to moderate and then difficult. Therefore, conditions' efficiency is evaluated accordingly. The total mean time users needed to perform recall and recognition tasks is presented in Figure 4-7. The highest time was recorded while performing the task related to the

NMMC (control condition) and the lowest time was recorded in the avatar condition. Overall, users spent more time on performing recall tasks than on recognition tasks.

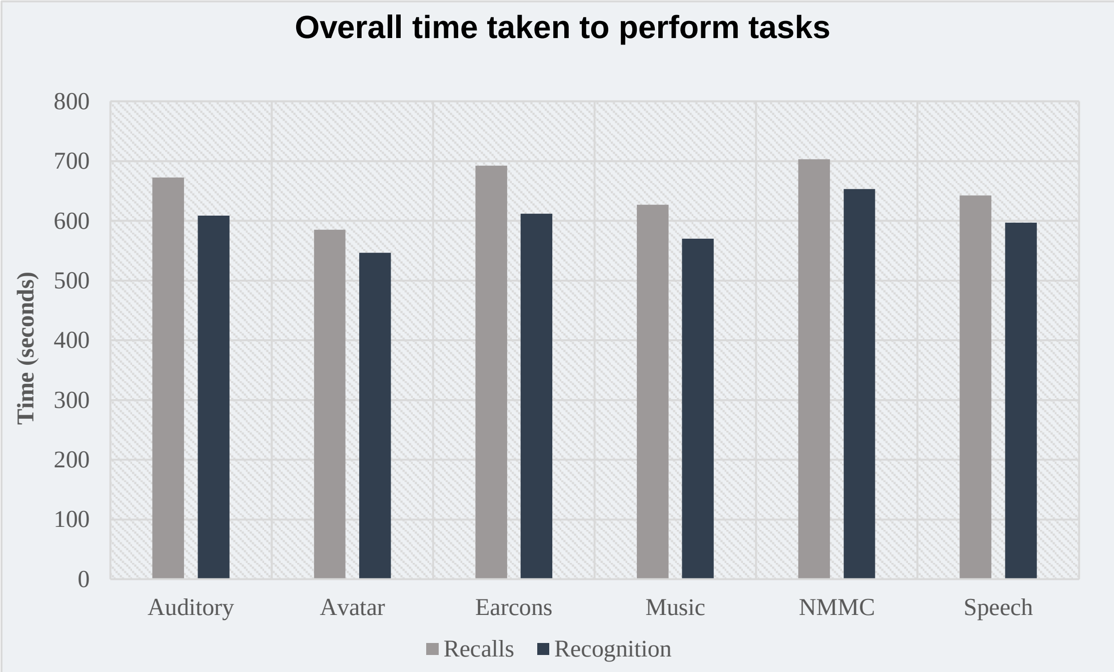


Figure 4-7: Time taken to perform recall and recognition tasks.

Efficiency of Recall Tasks: Concerning time taken to perform recall tasks, it seems that this increased with the increase in task complexity. Overall, on average users performed the avatar condition related recall tasks in 19.50 seconds, followed by 20.91seconds for the music condition, 21.41 seconds for the speech condition, 22.41 seconds for the auditory icons condition and 23.09 seconds for the earcons condition which was the highest time recorded in the experimental conditions. The highest time, 24.67 seconds, was recorded in the control condition (NMMC) and it seems that the inclusion of communication metaphors lowered the time needed to perform the tasks presented and increased the conditions' efficiency (see Figure 4-8).

Efficiency of Recognition Tasks: In relation to the efficiency of conditions in performing recognition tasks, the overall trend in Figure 4-8 shows clearly that the time spent by users increased in line with the escalation in tasks' complexity. Overall, the control condition (NMMC) in performing recognition tasks took the longest time at 23.66 seconds, followed by 22.25 seconds in the earcons condition, 20.69 seconds in the speech condition, 19.30 seconds in the auditory condition, 19.00 seconds in the music condition and the lowest time 18.22 seconds was taken in performing the avatar condition related tasks.

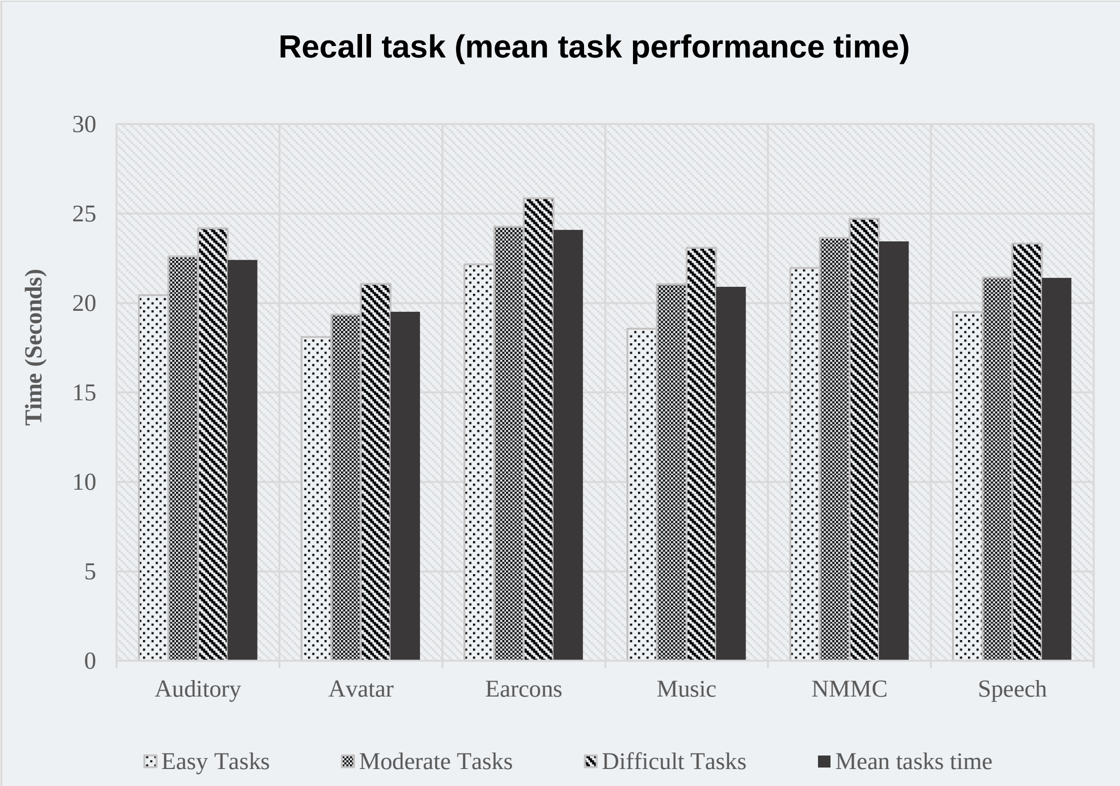


Figure 4-8: The efficiency recall tasks.

Figure 4-10, shows a comparison between overall task performance times in relation to task type; the general trend clearly indicates that the task performance time increased consistently with the increase in cognition load as performing recall tasks increased the time needed to perform the tasks. Tasks in the avatar condition were performed the fastest at 18.86 seconds followed by those in the music condition in 19.95 seconds, then 20.73 seconds in the speech condition and 21.36 seconds in the auditory condition. The longest times were recorded in the earcons condition and the NMMC at 23.17 seconds and 23.73 seconds respectively.

It is worth mentioning here that the difference between users' speed in performance recall tasks was less noticeable in the avatar and the speech conditions. It seems that the inclusion of avatar and speech communication metaphors improved users' ability to perform tasks with increased recognition loads (see Figure 4-9).

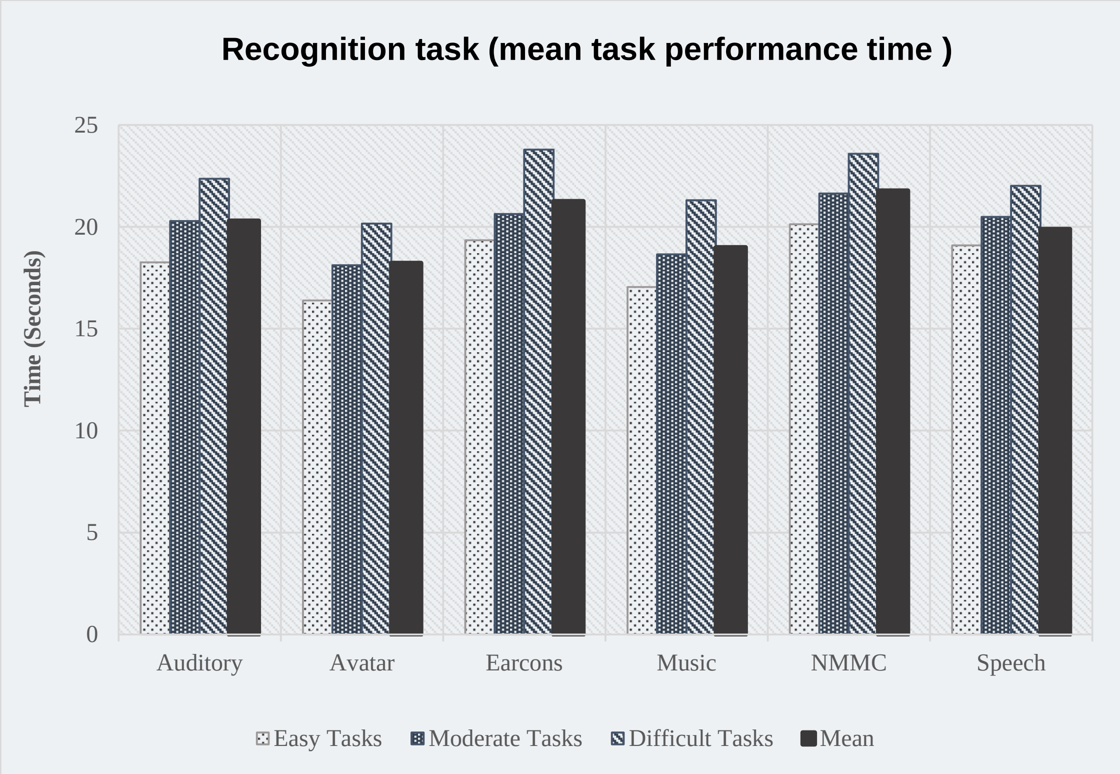


Figure 4-9: Recognition tasks efficiency.

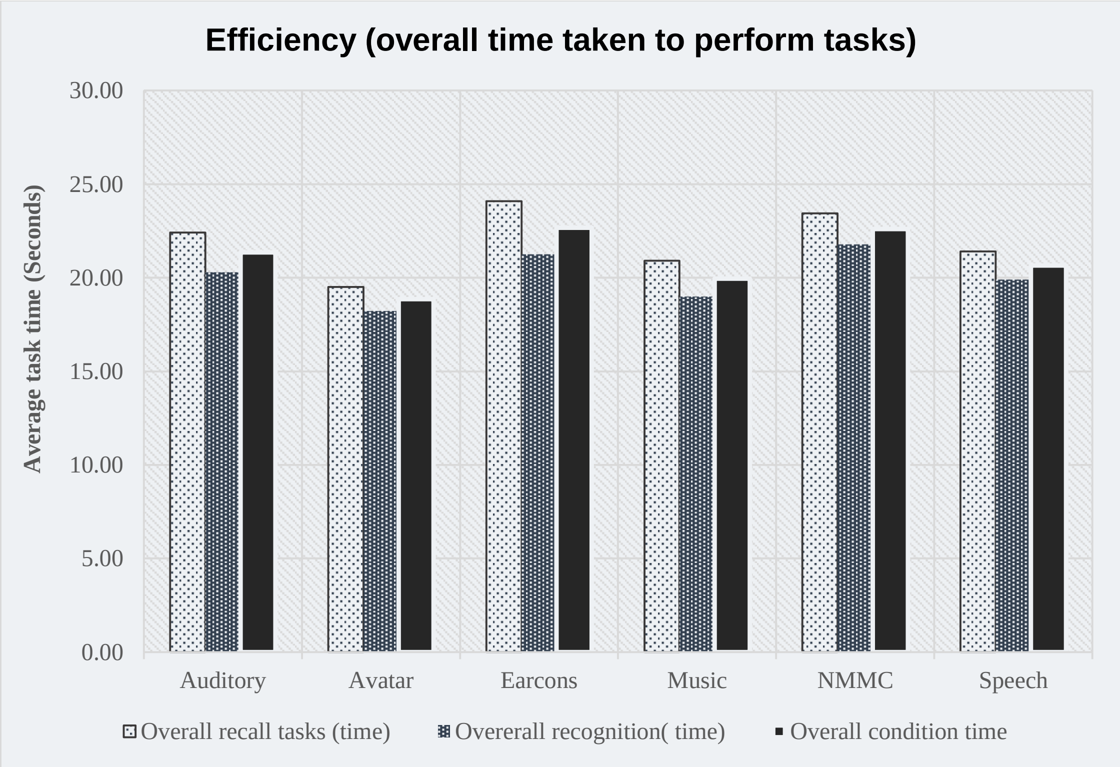


Figure 4-10: Overall tasks performance time in relation to their type.

4.6.3.2 *Statistical evaluation*

To examine whether there is statistical significance between the experimented conditions in terms of task performance efficiency, the overall task performance times between conditions will be examined using SPSS2013.

To determine the proper statistical test there is a need to examine whether the data is normally distributed using the Shapiro-Wilk normality test, as the sample size is less than 50, in addition to boxplots to determine whether there are outliers in the data.

Normality Test: The results of the conducted Shapiro-Wilk normality test in Table 4-2 clearly indicate that the data is normally distributed for all conditions ($p>0.05$). Since the experimental design uses within-users repeated measurements and the data is normally distributed, one-way repeated measures analysis of variance (ANOVA) test is the recommended test to be used to measure any differences in task performance efficiency within tasks.

Tests of normality						
	Kolmogorov-Smirnov[a]			**Shapiro-Wilk**		
Condition	Statistic	df	Sig.	**Statistic**	**df**	**Sig.**
Auditory	.089	30	.200*	.971	30	.573
Avatar	.126	30	.200*	.978	30	.767
Earcons	.116	30	.200*	.965	30	.420
Music	.068	30	.200*	.981	30	.839
Speech	.112	30	.200*	.967	30	.454
NMMC	.110	30	.200*	.975	30	.677
***. This is a lower bound of the true significance.**						
a. Lilliefors Significance Correction						

Table 4-2: Tasks efficiency normality test.

To use a one-way repeated measures ANOVA test the data have to meet the following assumptions:

- ✓ Have one dependent variable.
- ✓ Have three or more within-subjects independent variables.

- ✓ Have no significant outliers in the data.
- ✓ Be normally distributed.
- ✓ Have sphericity (the variances of the differences between all combinations of levels of the within-subjects variables must be equal).

All the above-mentioned assumptions were met, consequently to test the data for sphericity assumption, “Mauchly's” test of sphericity is used and the result is presented in Table 4-3. The test result shows that sphericity assumption has not been violated ($\chi2$ (14) = 18.443, p = .189).

Mauchly's Test of Sphericity[a]			
Measure: Tasks’ Efficiency			
Within Subjects Effect	**Mauchly's W**	**Approx. Chi-Square**	**Sig.**
Effectiveness	.506	18.443	.189
Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.			
a. Design: Intercept Within Subjects Design: Efficiency			

Table 4-3: Mauchly's test of sphericity.

Consequently, it is appropriate to carry on using the one-way repeated measures ANOVA test.

One-Way Repeated Measures ANOVA Test

There is a null hypothesis (H0) and an alternative hypothesis (HA) to test; both are stated below.

$$\text{H0: } \mu_{\text{Auditory}} = \mu_{\text{Avatar}} = \mu_{\text{Earcons}} = \mu_{\text{Music}} = \mu_{\text{NMMC}} = \mu_{\text{Speech}}.$$

$$\text{HA: } \mu_{\text{Auditory}} \neq \mu_{\text{Avatar}} \neq \mu_{\text{Earcons}} \neq \mu_{\text{Music}} \neq \mu_{\text{NMMC}} \neq \mu_{\text{Speech}}.$$

The one-way repeated measures ANOVA test was conducted and the full test results presented in Table 4-4.

The test result indicated that, the difference (F (5, 145) = 125.481, p < .001) between the tested conditions in terms of their effectiveness in performing the presented tasks is statistically significant.

Accordingly, the H0 is rejected and HA is accepted. Unfortunately, the one-way repeated measures ANOVA test did not show exactly where the difference lies or which condition was more effective.

Tests of Within-Subjects Effects							
Measure: Task performance effectiveness							
Source	**Type III Sum of Squares**		**df**	**Mean Square**	**F**	**Sig.**	**Partial Eta Squared**
Task performance time	Sphericity Assumed	300.674	5	60.135	125.48	.000	.812
Error(Task performance time)	Sphericity Assumed	69.489	145	.479			

Table 4-4: One-way repeated measures ANOVA test result.

To find which experimental condition contributed more to the perceived difference in the obtained result, the Bonferroni post hoc test is recommended by Maxwell & Delaney [202]. It tests the statistical significance difference in all possible pairwise combinations within-subjects measurements and provides confidence intervals for the mean difference for each pairwise. The *post-hoc* test results are presented in Table 4-5. For presentation reasons duplicated pairs were removed from the table.

Overall, the difference between all conditions reached a statistically significant level except for two conditions, the NMMC (control condition) and the earcons condition. The results indicated that all communication metaphors improved task performance efficiency except for the earcons communication metaphor. Figure 4-12 presents, in the form of a graph, the means ranks of the tested conditions according to their efficiency. The most effective metaphor was the avatar, followed by the music, speech, and auditory icons conditions. The least effective communication metaphor was the earcons.

Tasks efficiency post-hoc test						
(I) Effectiveness	(J) Effectiveness	Mean Difference (I-J)	Std. Error	Sig.[b]	95% Confidence Interval for Difference[b]	
					Lower Bound	Upper Bound
Auditory	Avatar	2.492*	.225	.001	1.773	3.212
	Earcons	-.888*	.189	.001	-1.492	-.283
	Music	1.402*	.230	.001	.667	2.137
	Speech	.701*	.191	.014	.091	1.311
	NMMC	-1.253*	.199	.001	-1.889	-.617
Avatar	Earcons	-3.380*	.142	.001	-3.834	-2.926
	Music	-1.090*	.156	.001	-1.589	-.592
	Speech	-1.791*	.174	.001	-2.348	-1.234
	NMMC	-3.745*	.197	.001	-4.375	-3.116
Earcons	Music	2.290*	.133	.001	1.864	2.715
	Speech	1.589*	.148	.001	1.115	2.062
	NMMC	-.365	.157	.412	-.869	.138
Music	Speech	-.701*	.149	.001	-1.178	-.224
	NMMC	-2.655*	.183	.001	-3.242	-2.068
Speech	NMMC	-1.954*	.174	.001	-2.512	-1.396
Based on estimated marginal means						
*. The mean difference is significant at the .05 level.						
b. Adjustment for multiple comparisons: Bonferroni.						

Table 4-5: *Post-hoc* test (pairwise comparisons) results.

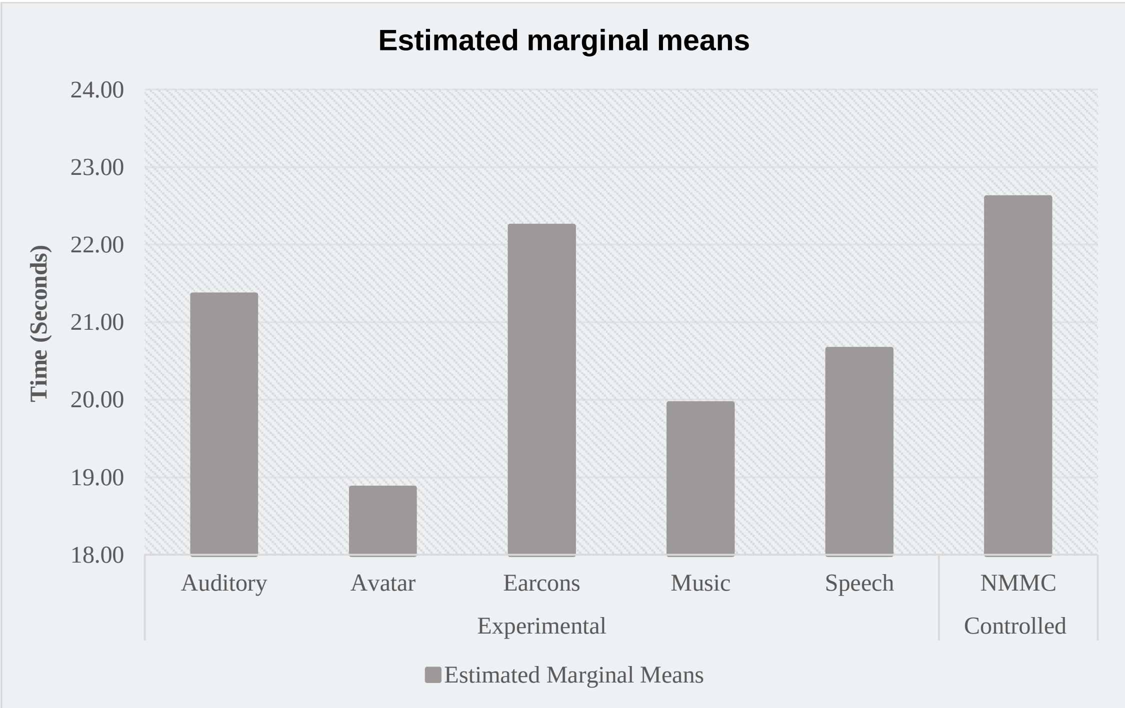


Figure 4-11: Estimated marginal means of task performance time.

4.6.3.3 *Effectiveness*

Effectiveness refers to the correctness of performed tasks. Users were presented with two types of task, recall, and recognition. Both types of task were designed to have three levels of complexity ranging from easy to moderate and then difficult. The total score in each condition and the total scores for each task type and complexity level are depicted in Figure 4-12. Concerning the total scores for all types of task and complexity levels, the avatar condition achieved the highest score with 159 points while the lowest score was achieved in the earcons condition with 75 points (see Figure 4-14).

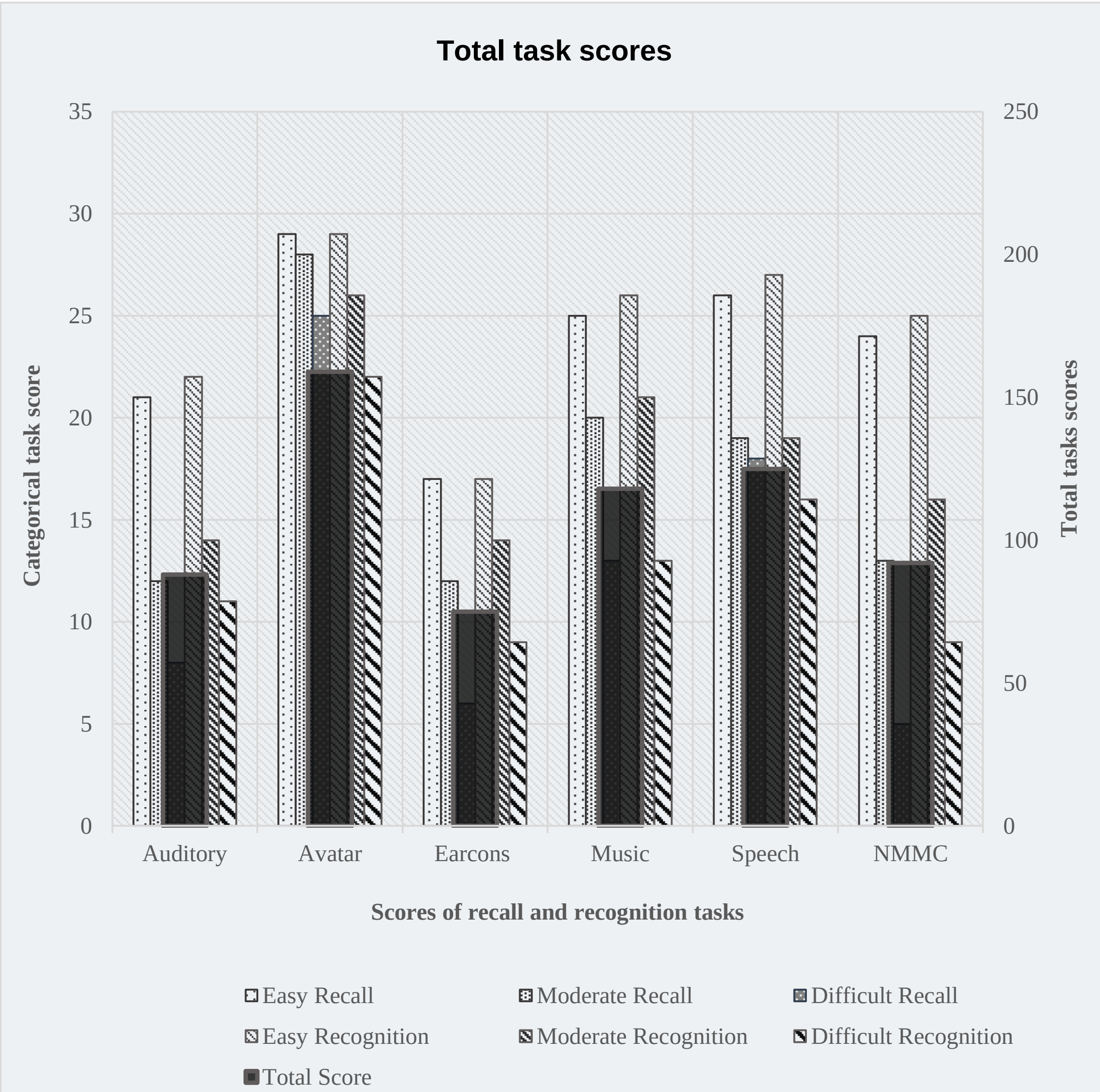


Figure 4-12: Total scores for recall and recognition tasks.

Recall Tasks Effectiveness: The results of the recall tasks are presented in Figure 4-14. Overall, the results can be categorized into two groups. In the first group, the effectiveness of conditions was better than the control condition (NMMC, the mean score was 52%), where on average the avatar condition score was 90%, followed by 70% for the speech condition and 63% for the music condition. The task effectiveness in the second group was less than that for the experimental condition with 47% for the auditory condition and only 40% for the earcons condition. With regard to task complexity, it is noted that conditions' effectiveness decreased in line with the increase in task complexity; however, the decrease is less severe in the avatar condition. Figure 4-14 shows the effectiveness of conditions in performing recognition tasks. Overall, the least effective condition was the earcons condition which scored 46%, followed by the auditory condition with 52%.

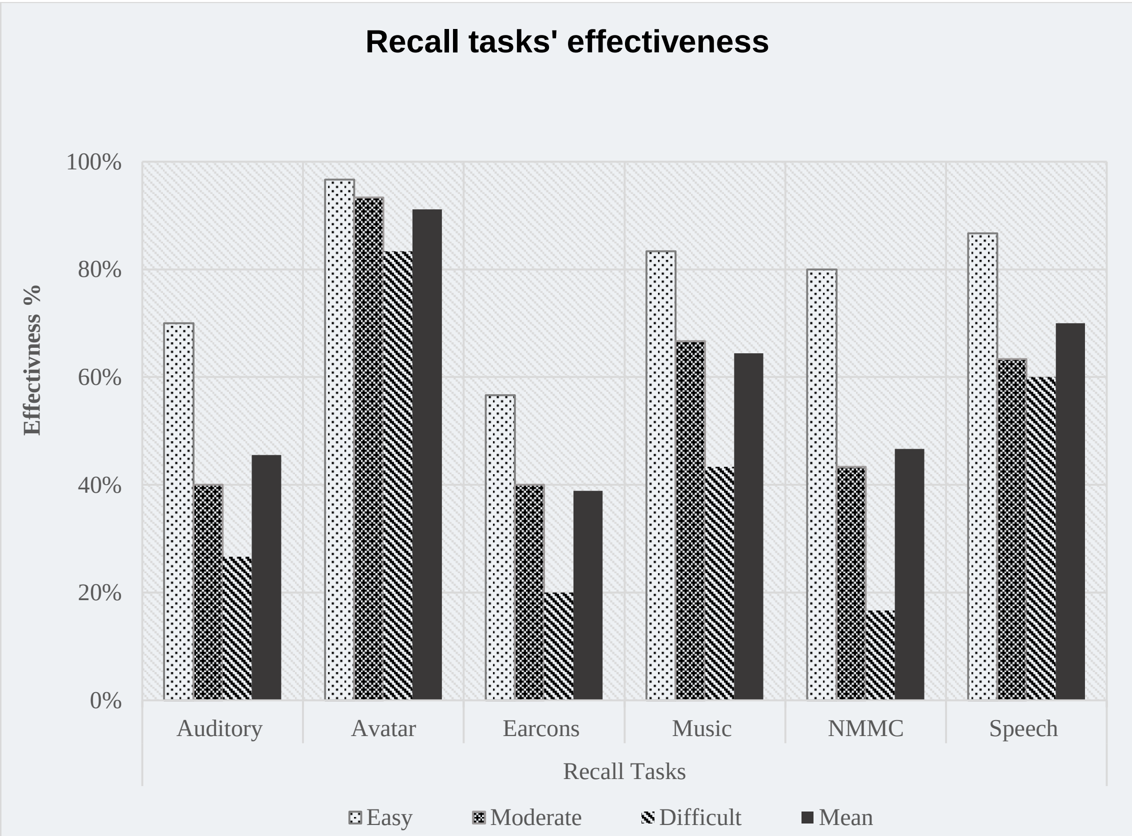


Figure 4-13: Conditions effectiveness in performing recall tasks.

Both conditions were less effective than the control condition, which scored 56%. The remaining conditions were more effective in performing recognition tasks, with the avatar condition scoring the highest with a mean value of 86%, followed by 69% in the speech condition and 67% in the music condition. Concerning task complexity, the decline in effectiveness was consistent with increment of task complexity.

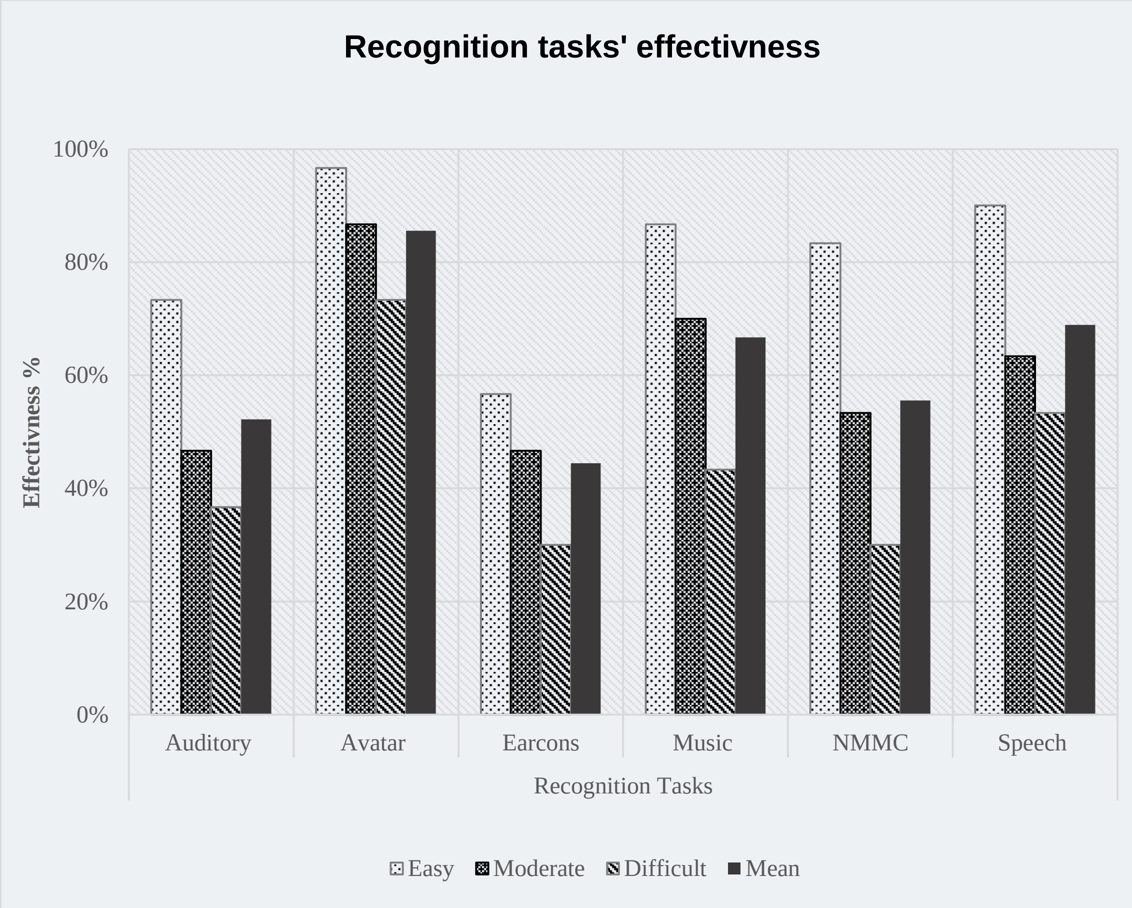


Figure 4-14: Condition effectiveness in performing recognition tasks.

Overall Task Effectiveness: Figure 4-15 shows the overall effectiveness of tasks in terms of experimental conditions that were more effective than the control condition (NMMC). The avatar condition was the most effective with a score of 73.33%, followed by 53.33% in the speech condition, then 43.33% in the music condition and 36.66% in the auditory condition. There was no improvement in the effectiveness of the earcons condition in comparison with the NMMC; both conditions were effectively equal with a score of 30%. Regarding condition effectiveness in relation to task types and cognition load, from the difference line graph in Figure 4-15 it seems that the increase in task cognition load had a negative impact on condition effectiveness except for the avatar and speech conditions where condition effectiveness was maintained in line with task cognition load and complexity.

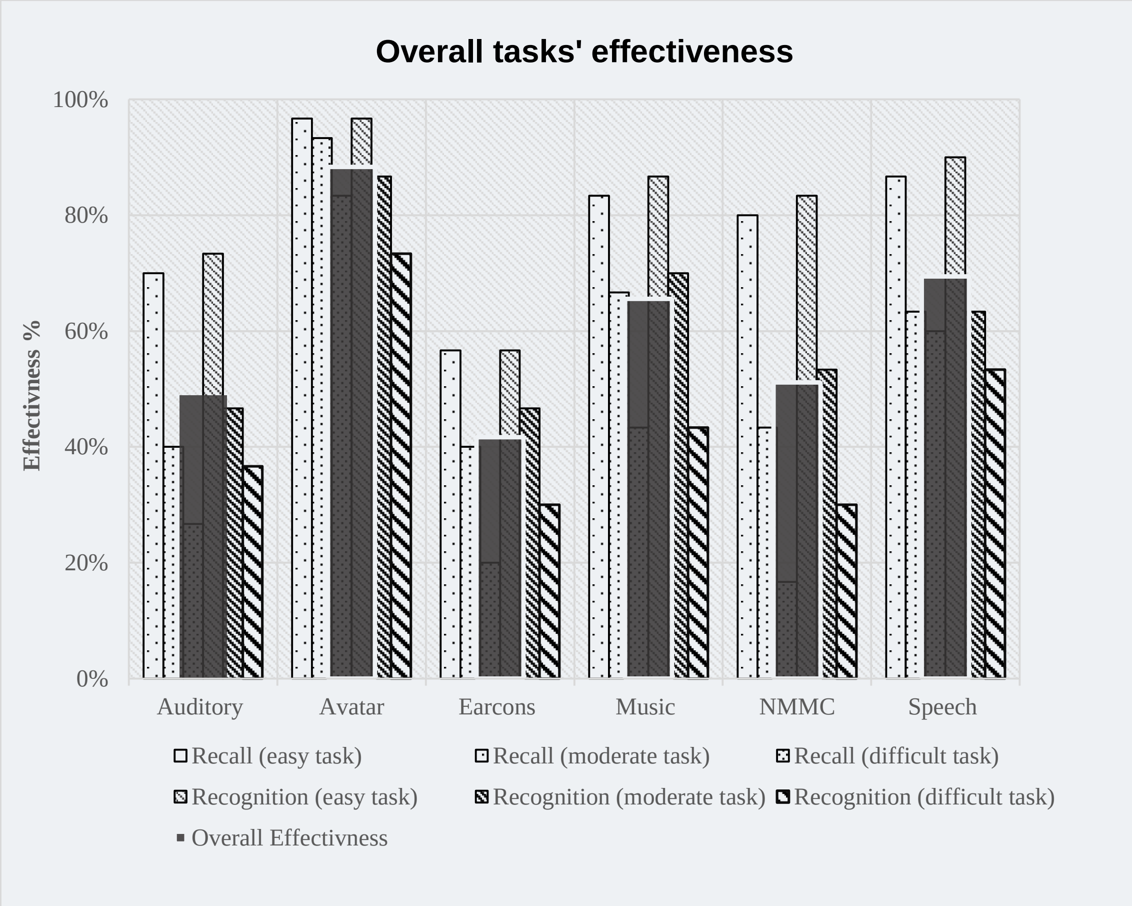


Figure 4-15: Overall tasks' effectiveness.

4.6.3.4 *Statistical evaluation*

To examine whether there is statistical significance in the difference between the experimented conditions in terms of task performance efficiency, the overall task performance times between conditions will be examined using SPSS. To determine the proper statistical test, there is a need to examine whether the data is normally distributed.

Normality test: Due to the sample size which is less than 50 , the Shapiro-Wilk normality test was used , the results in Table 4-6 plainly point out that the data is not normally distributed for all conditions ($p<0.05$). Consequently, to test the difference between conditions in terms of effectiveness, a non-parametric statistical test (related samples Friedman’s analysis of variance (ANOVA)) was used. The results of the test and the means rank are presented in Figure 4-16.

Effectiveness normality test						
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
Condition	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.207	30	.002	.891	30	.005
Avatar	.284	30	.000	.774	30	.000
Earcons	.234	30	.000	.871	30	.002
Music	.378	30	.000	.720	30	.001
Speech	.385	30	.000	.704	30	.001
NMMC	.345	30	.000	.750	30	.001
a. Lilliefors Significance Correction						

Table 4-6: Normality test of effectiveness.

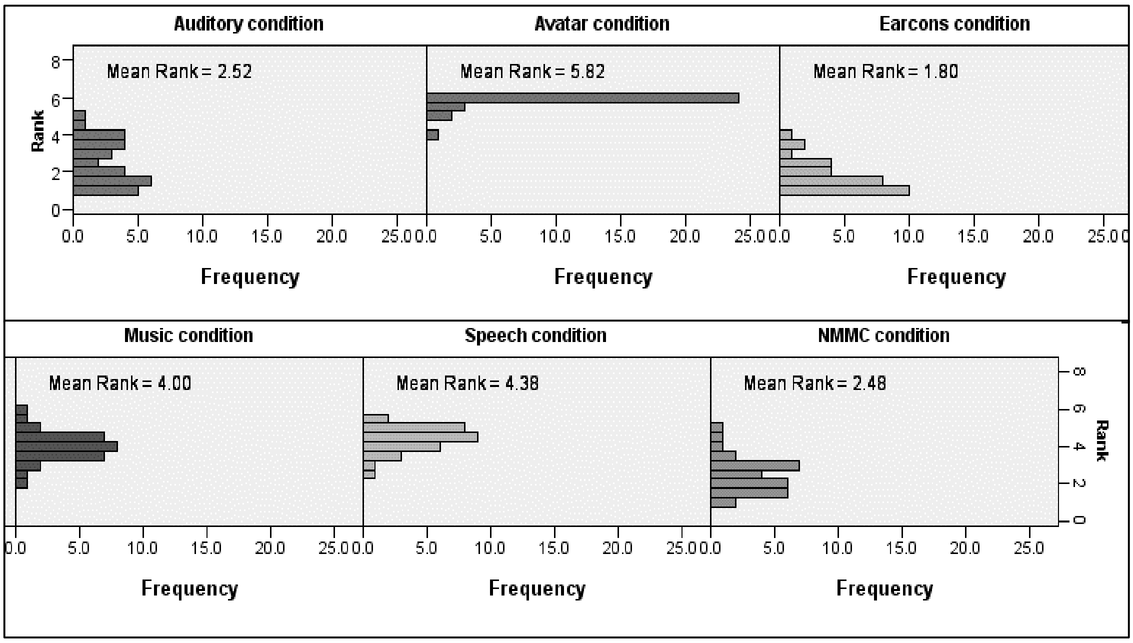
Effectiveness Related Samples Friedman's ANOVA Test: There is a null hypothesis (H0) and an alternative hypothesis (HA) to test and they are stated below.

$$\text{H0: } \mu_{\text{Auditory}} = \mu_{\text{Avatar}} = \mu_{\text{Earcons}} = \mu_{\text{Music}} = \mu_{\text{NMMC}} = \mu_{\text{Speech}}$$

$$\text{HA: } \mu_{\text{Auditory}} \neq \mu_{\text{Avatar}} \neq \mu_{\text{Earcons}} \neq \mu_{\text{Music}} \neq \mu_{\text{NMMC}} \neq \mu_{\text{Speech}}$$

The Friedman's test result ($\chi 2$ (5) = 107.960, $p < .0005$.) showed a difference between conditions in their effectiveness. Therefore, the null hypothesis is rejected and the alternative hypothesis is accepted. A *post-hoc* test is needed to define the source of the variance. Pairwise comparisons were performed (using SPSS2013) with a Bonferroni correction for multiple comparisons (to avoid making type I error, which is very common in this scenario) and the results are presented in Figure 4-17 (gray denotes a statistically significant pairwise comparison). The *post-hoc* analysis revealed statistically significant differences in conditions' effectiveness on three levels. The highest significance level was in the avatar condition (Mdn=83%, $p<0.05$) in comparison with the control condition (NMMC) and the rest of the conditions (with five significances). The second level includes the music (Mdn=67%, $p<0.05$) and speech (Mdn=67%, $p<0.05$) conditions (with four significances) which were significantly different to the rest of the conditions; however there is no significant difference between the condition within the same level (Mdn=67 ,$p>0.05$. The last level includes the auditory condition (Mdn=50% ,$p>0.05$) as well as the earcons condition (Mdn=41%, p>0.05) and the NNMC (Mdn=50%, p>0.05). The three conditions of avatar, music and speech were effectively alike, however they significantly differed from the other two conditions (for graphical illustration please refer to Figure 4-17 in which significant differences are denoted in gray). With reference to pairwise comparison results, in addition to mean and medians rank, the avatar condition was the most effective, followed by the speech condition, then the music condition; these all performed better than the control condition (NMMC). The least effective condition was the earcons condition, followed by the auditory condition; these were both less effective than the control condition. Figure 4-18 ranks conditions according to their effectiveness.

Effectiveness related samples Friedman's ANOVA test



Hypothesis Test Summary

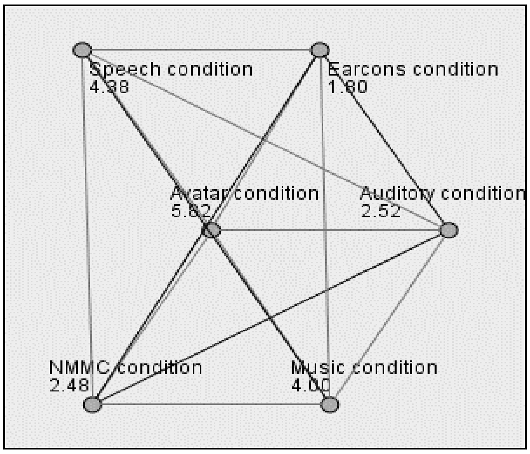
	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory condition, Avatar condition and Earcons condition are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	107.960
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 4-16: Related samples Friedman's ANOVA test result.

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Earcons condition-NMMC condition	-.683	.483	-1.415	.157	1.000
Earcons condition-Auditory condition	.717	.483	1.484	.138	1.000
Earcons condition-Music condition	-2.200	.483	-4.554	.000	.000
Earcons condition-Speech condition	-2.583	.483	-5.348	.000	.000
Earcons condition-Avatar condition	4.017	.483	8.315	.000	.000
NMMC condition-Auditory condition	.033	.483	.069	.945	1.000
NMMC condition-Music condition	1.517	.483	3.140	.002	.025
NMMC condition-Speech condition	1.900	.483	3.933	.000	.001
NMMC condition-Avatar condition	3.333	.483	6.901	.000	.000
Auditory condition-Music condition	-1.483	.483	-3.071	.002	.032
Auditory condition-Speech condition	-1.867	.483	-3.864	.000	.002
Auditory condition-Avatar condition	-3.300	.483	-6.832	.000	.000
Music condition-Speech condition	-.383	.483	-.794	.427	1.000
Music condition-Avatar condition	1.817	.483	3.761	.000	.003
Speech condition-Avatar condition	1.433	.483	2.967	.003	.045

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 4-17: Effectiveness pairwise comparison.

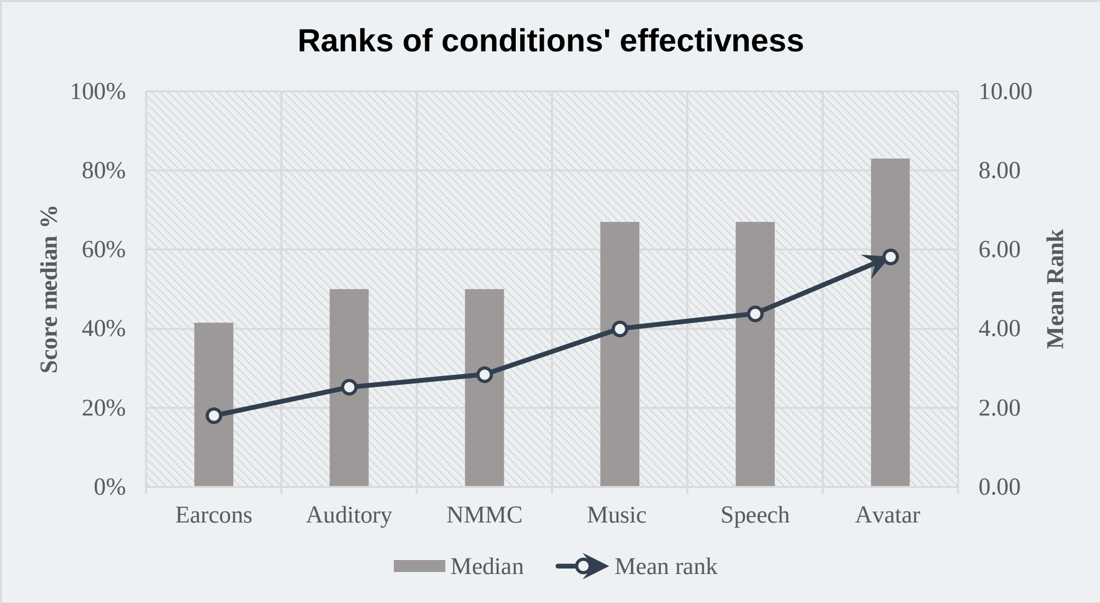


Figure 4-18: Efficiency ranks.

4.6.3.5 *Learning Performance*

Learning is a task that can be evaluated within two dimensions, learning time and learning outcomes. In relation to this study, these two dimensions are the actual learning time and task scores. Figure 4-19 shows the overall learning time and task scores for each tested condition.

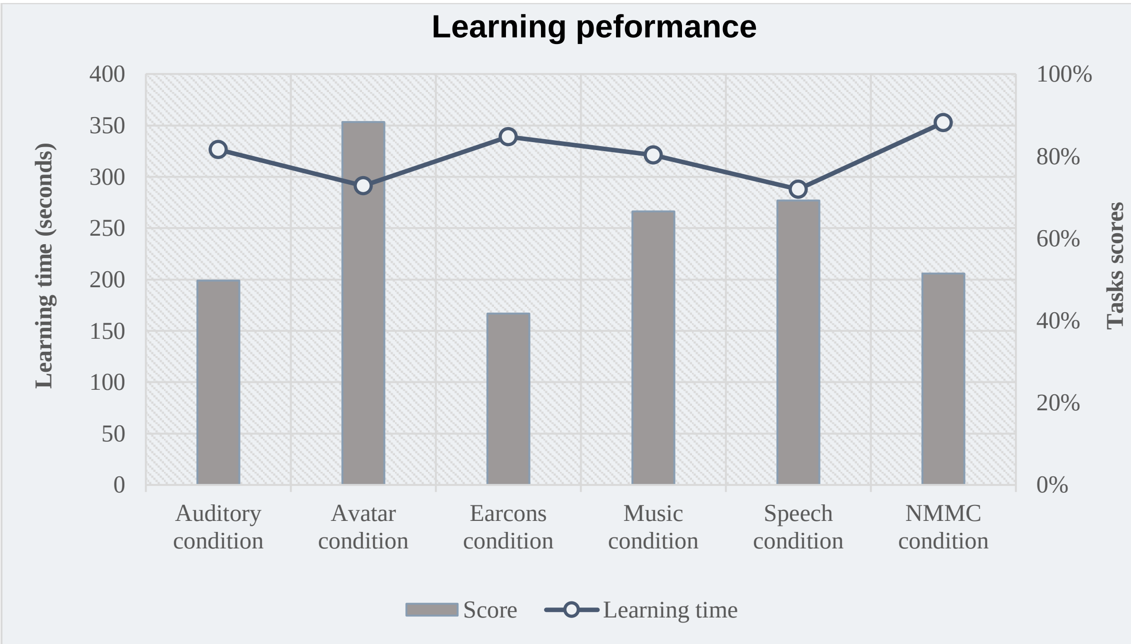


Figure 4-19: Overall learning time and task scores.

For a better understanding of the relation between learning time and learning outcome, a scatterplot was produced (see Figure 4-20).

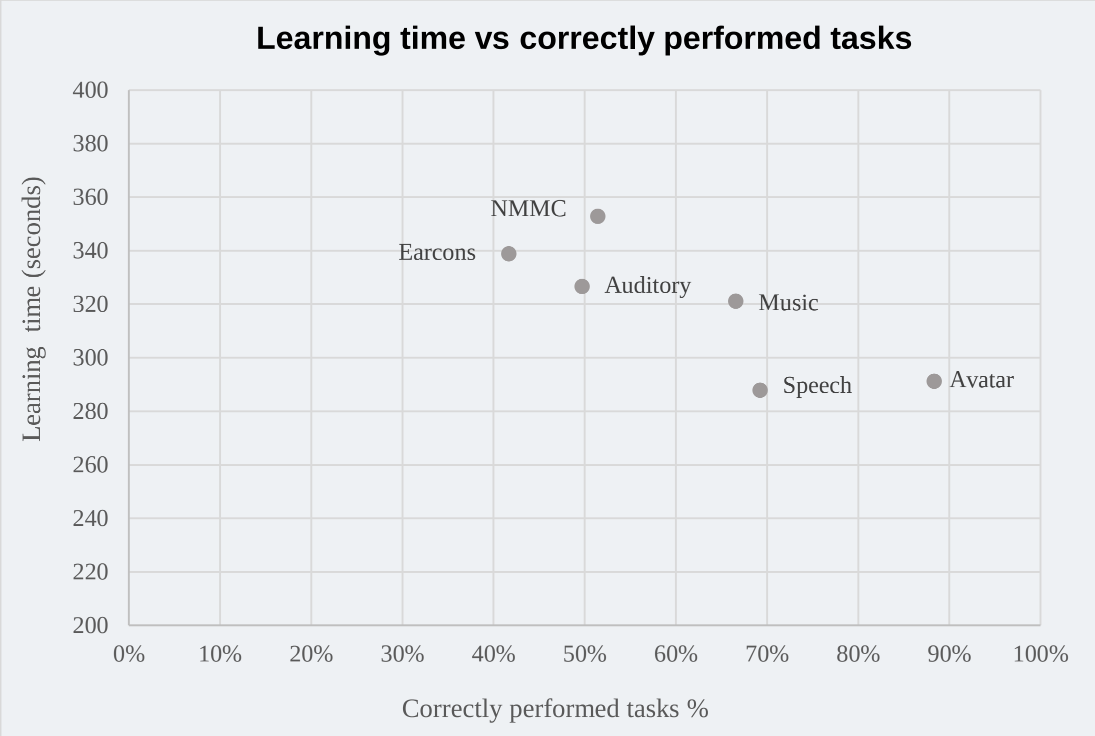


Figure 4-20: Learning performance (learning time vs learning outcome).

The figure shows that three conditions outperformed the NMMC (score = 51%, learning time = 352.83s) in learning performance; in descending order these are avatar condition (score = 88%, learning time = 291.23s), followed by speech condition (score=69%, learning time = 287.87s) and music condition (score = 67.32%, learning time = 321.09s). The conditions, which did not perform as well as the control condition, were the auditory condition (score=50%, learning time = 326.60s) followed by the earcons condition (score=69%, learning time = 287.87s) which was the least effective of all experimented conditions. To find the association between learning time and learning outcomes in relation to learning performance and condition type, the following two hypotheses were formulated:

- Null hypothesis (H0): There is no correlation between learning time and outcomes.
- Alternative hypothesis (HA): There is a correlation between learning time and outcomes.

Pearson correlation analysis (SPSS 2013) was used to obtain the correlation coefficient and significance level and the results are listed in Table 4-7. There was a significant negative correlation between learning time and outcomes in learning performance (r = -.823, ρ = 0.044) within the tested conditions. Consequently the HA is accepted and the H0 is rejected. The results indicate that increase in learning time did not result in improved outcomes, in fact the observed improvements can only be attributed to the use of the communication metaphors.

Correlations		Score	Time
Score	Pearson Correlation	1	-.823*
	Sig. (2-tailed)		.044
	N	6	6
Time	Pearson Correlation	-.823*	1
	Sig. (2-tailed)	.044	
	N	6	6
*. Correlation is significant at the 0.05 level (2-tailed).			

Table 4-7: Pearson correlation analysis.

4.6.3.6 *Users' satisfaction*

After the completing each one of the conditions, users were requested to complete a survey containing eight statements to evaluate their satisfaction. Seven of the statements were common to all conditions and one was condition-tailored to reflect its unique metaphor.

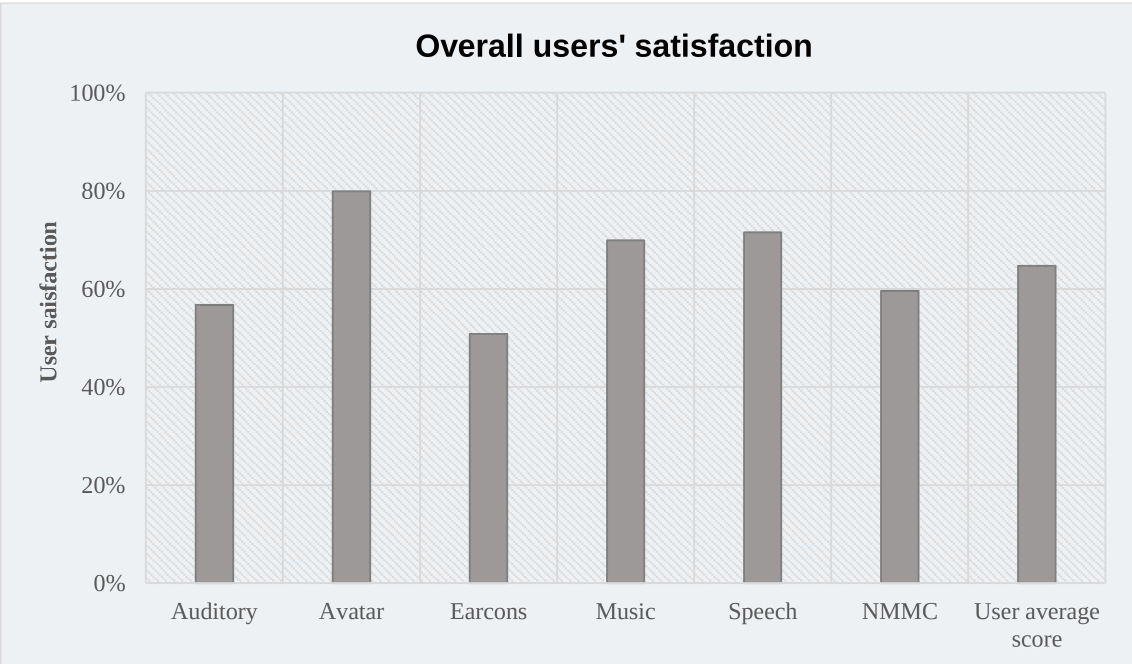


Figure 4-21: Overall users satisfaction.

Users were given a four-point Likert-scale to rate each one of the statements (strongly disagree, disagree, agree and strongly agree) and the overall users' satisfaction is presented in Figure 4-21. The satisfaction statements and their user ratings are presented below.

Ease of Use: Users were asked to rate the perceived ease of use concerning each one of the experimented conditions. They considered the avatar condition (rated=3.67) to be the easiest to use, followed by the speech condition (rated= 3.20), then the music and NMMC conditions respectively. The lowest rating was for the auditory condition (rated=2.63) followed by the earcons conditions (refer to Figure 4-22).

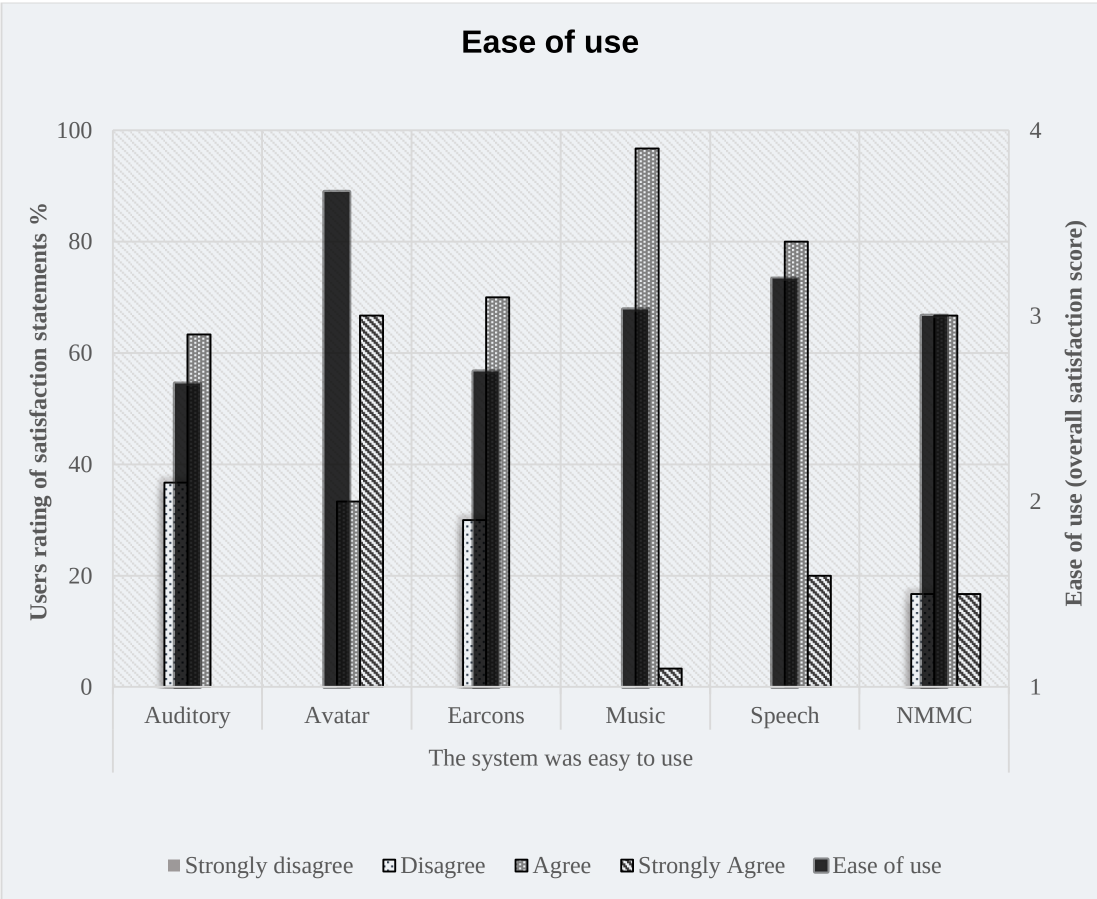


Figure 4-22: Ease of use.

Learnability: When users were asked to rate how easy it was to learn to use each one of the experimented conditions, they perceived the avatar condition as the easiest (rated = 3.5), the most difficult was the auditory condition (rated=2.53) followed by the NMMC condition and the rest were comparably equal (see Figure 4-23).

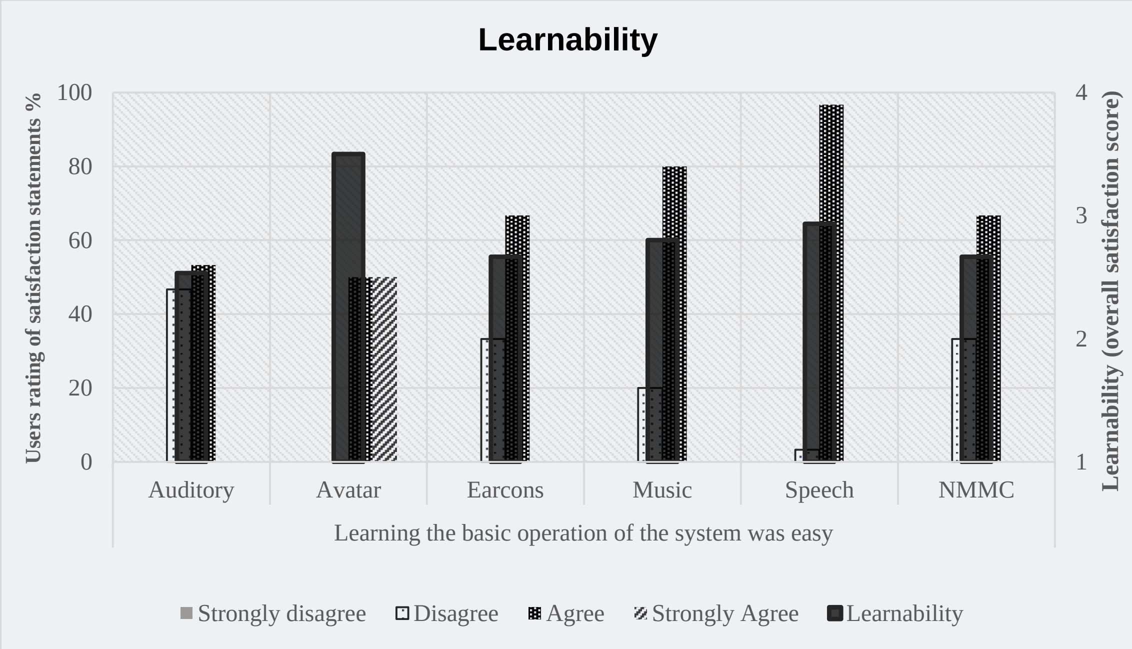


Figure 4-23: Learnability.

Convenience of use: When users were requested to rate how comfortable they were using each of the experimented conditions, they rated the avatar condition (rated=3.07) as the most convenient to use, followed by the music condition, then the speech and NMMC conditions respectively. The lowest rating was for the earcons condition (rated=1.47) followed by the auditory condition (refer to Figure 4-24).

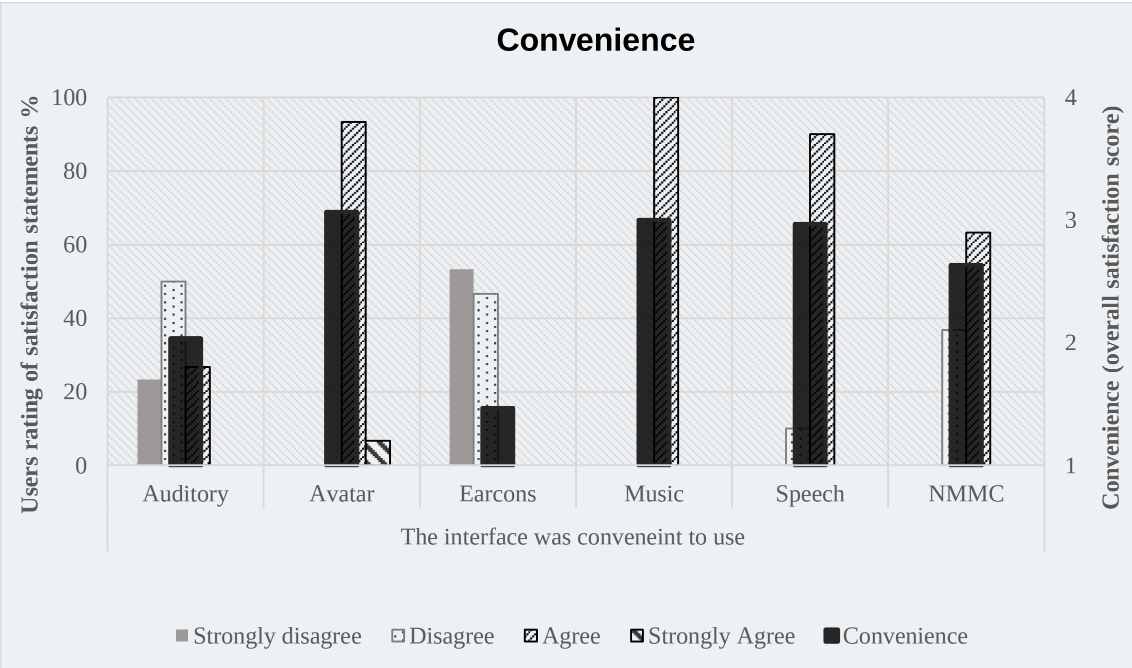


Figure 4-24: Convenience of use.

Likability: When users were asked to rate, their liking of each of the investigated conditions, they gave the avatar condition as the most liked condition (rated=3.30), followed by speech and music and the NMMC conditions respectively. The lowest rating was for the earcons condition (rated=1.67) followed by the auditory condition (refer to Figure 4-25).

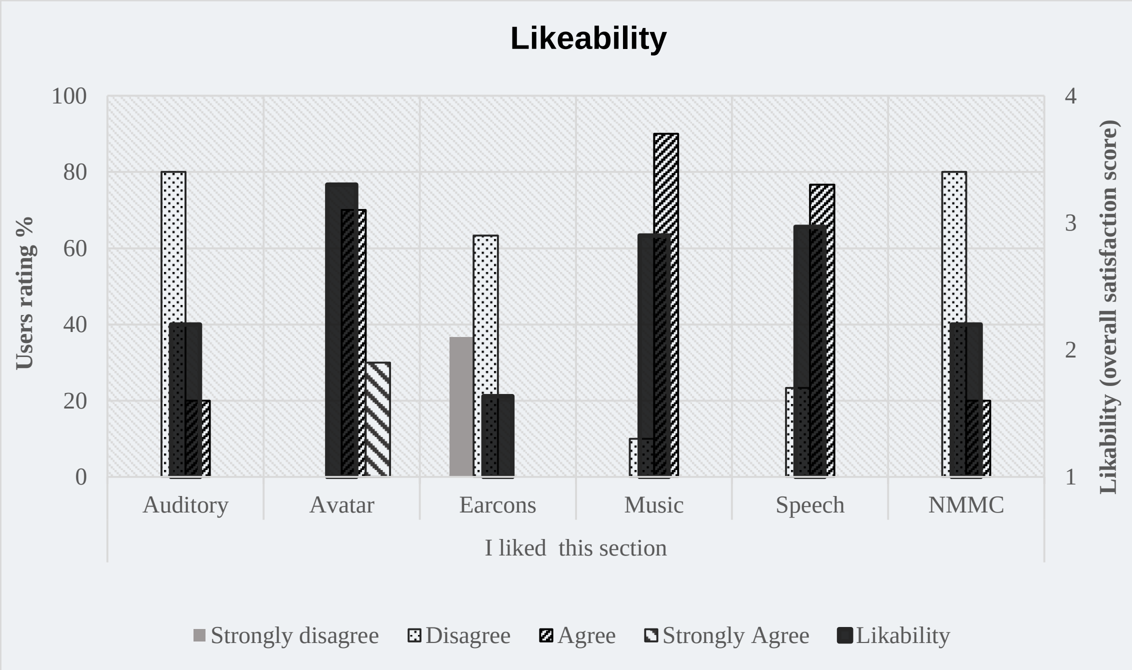


Figure 4-25: Likeability.

User engagement: When users were asked to value their engagement with each of the experimented conditions, users elected the avatar condition (rated=3.07) as the most engaging, followed by the speech and music conditions as they were relatively equally engaging. The next rank was for the NMMC condition (rated=2.27), trailed by the auditory condition and then the earcons condition (rated=1.50) which was seen as the least engaging condition (refer to Figure 4-26).

Human-to-human communications : When users were asked to rate to what degree they did not miss human-to-human communication while using each of the experimented conditions, the worst rating was in the NMMC condition (rated=1.37) followed by the auditory condition, then the earcons and music conditions respectively. Users' best rating was given for the avatar (rated=2.63) and the speech conditions correspondingly (refer to Figure 4-27, for a detailed view).

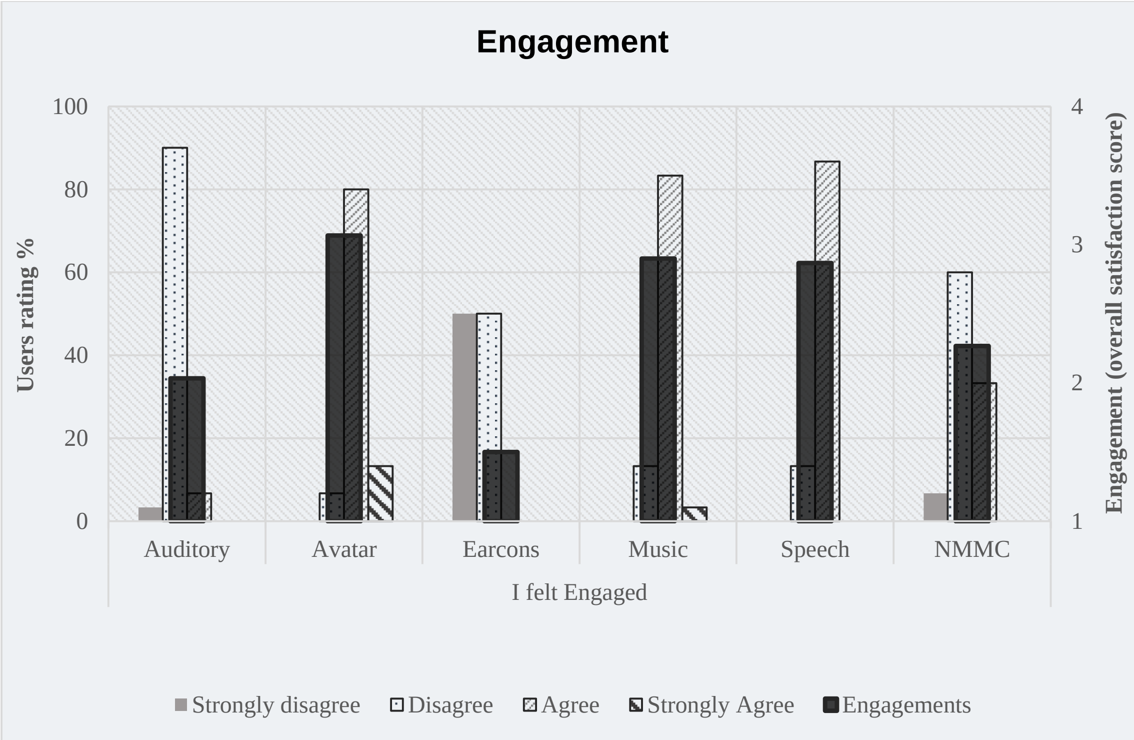


Figure 4-26: Users engagement.

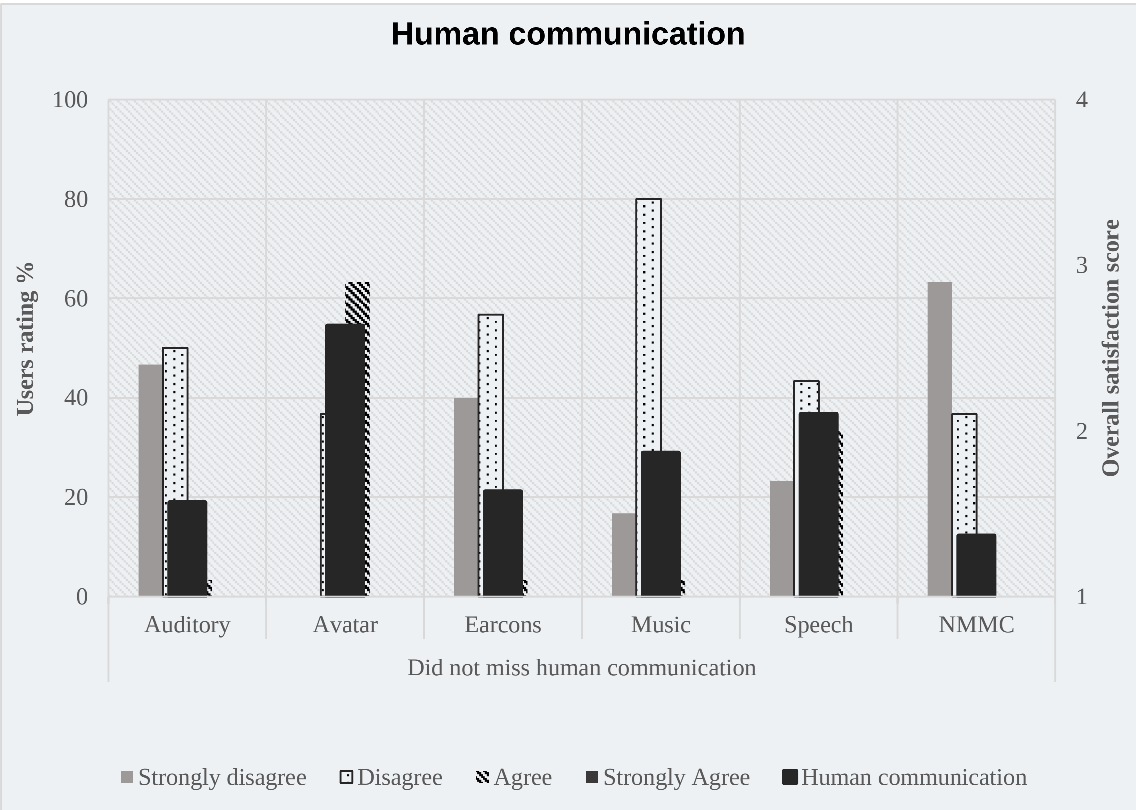


Figure 4-27: Human communication.

Satisfaction: When users were requested to rate how pleased they were with each one of the experimented conditions, they expressed more satisfaction (rated=3.03) with the avatar condition than any other, followed by the speech and music conditions which were considered equally satisfying. The next rank was for the auditory and NMMC conditions respectively. Out of all the tested conditions, users voted the earcons condition (rated=1.83) as the least satisfying (refer to Figure 4-28).

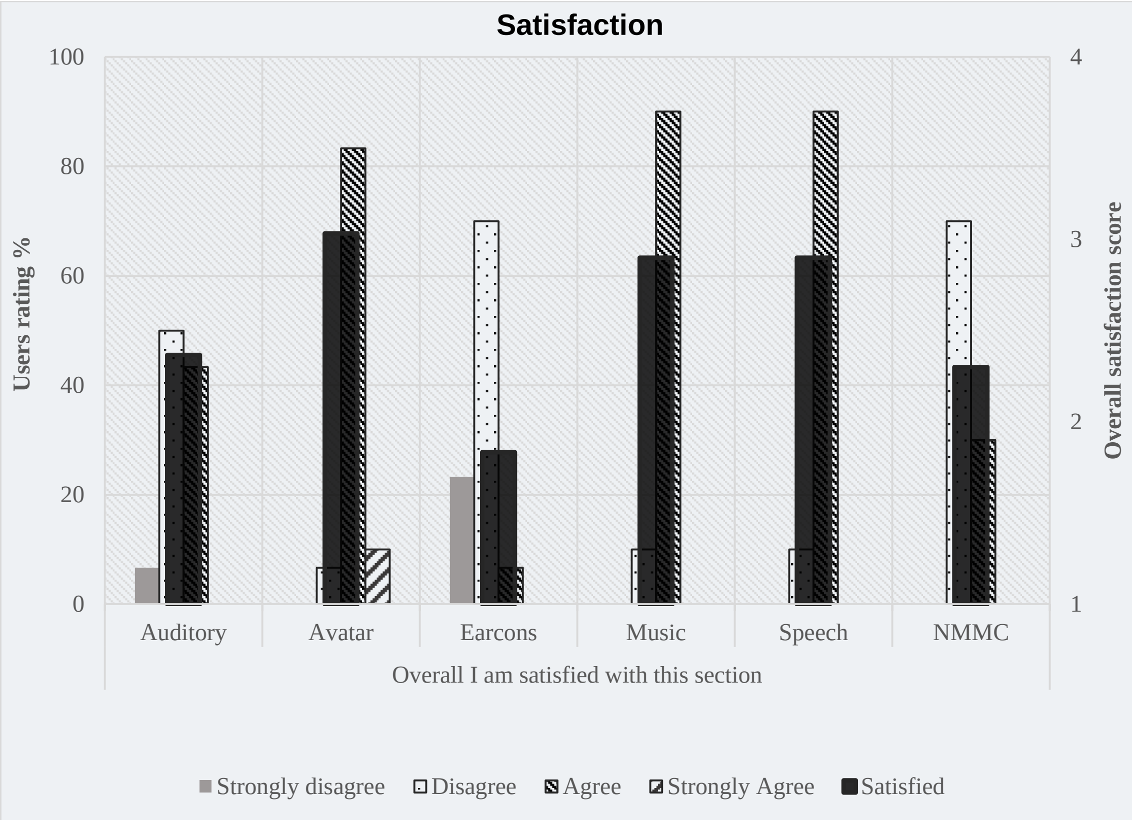


Figure 4-28: Satisfaction.

Condition specific statement: One of the satisfaction statements rated by users was specifically formulated in relation to the utilized communication metaphor (see Figure 4-29).

Users were relatively high in agreement with statements concerning all conditions; the statements and their rating are listed below in ascending order:

- The NMMC statement (rated=2.63) “without sound I was able to concentrate”
- The auditory icons condition statement (rated=2.8) “the sound made the lesson more realistic”

- The earcons condition statement (rated=2.83) “the sound alerted me to the important points”
- The music condition statement (rated=2.97) “the music helped me to concentrate”
- The speech condition statement (rated=2.97) “without having to read texts, I was able to concentrate”
- The avatar condition statement (rated=3.3) “the virtual lecturer helped me to understand”

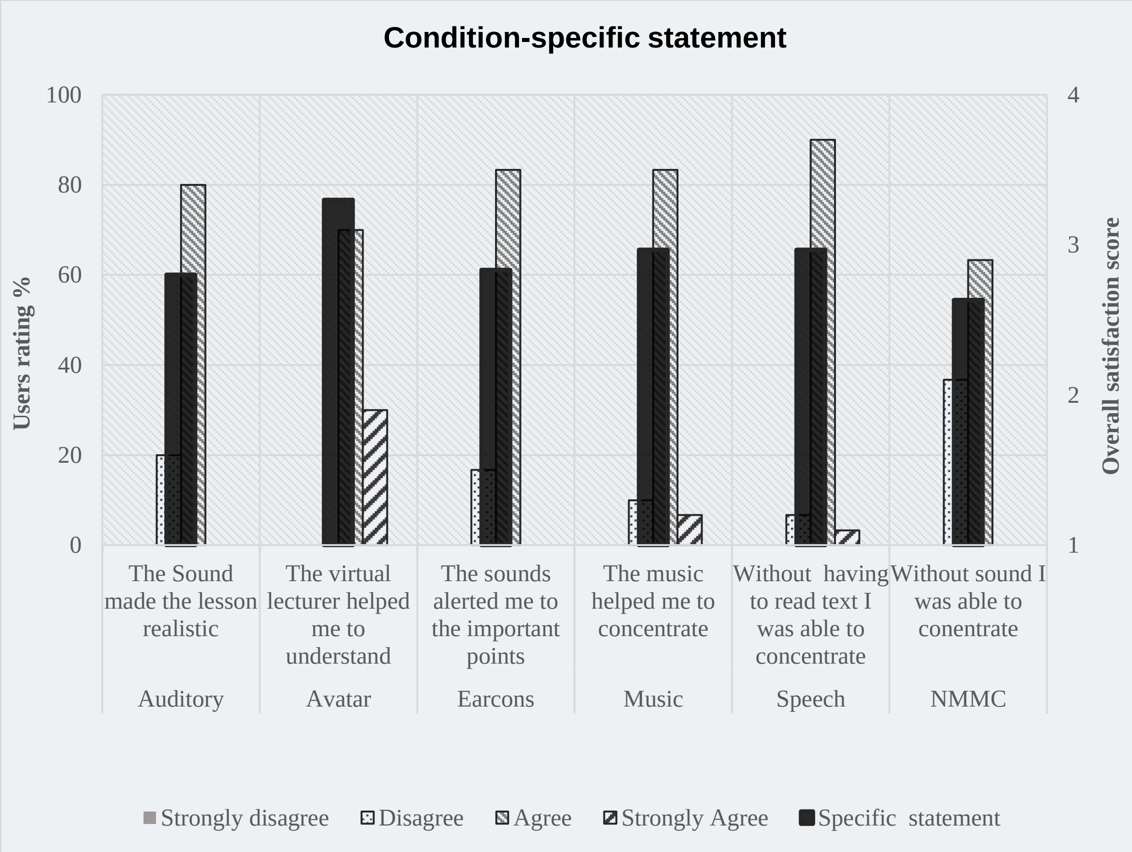


Figure 4-29: Condition specific statements.

Figure 4-30 presents the overall user ratings for experimented conditions in comparison with the control condition (NMMC); from the illustration, the results can be classified into two categories. The first category includes the conditions which outperformed the control condition (avatar, speech and music) and the second category contains those which underperformed (earcons and auditory). To find whether the differences in users’ satisfaction between the tested conditions reached a significant level a statistical data analysis was executed and the outcomes are presented in the following segments.

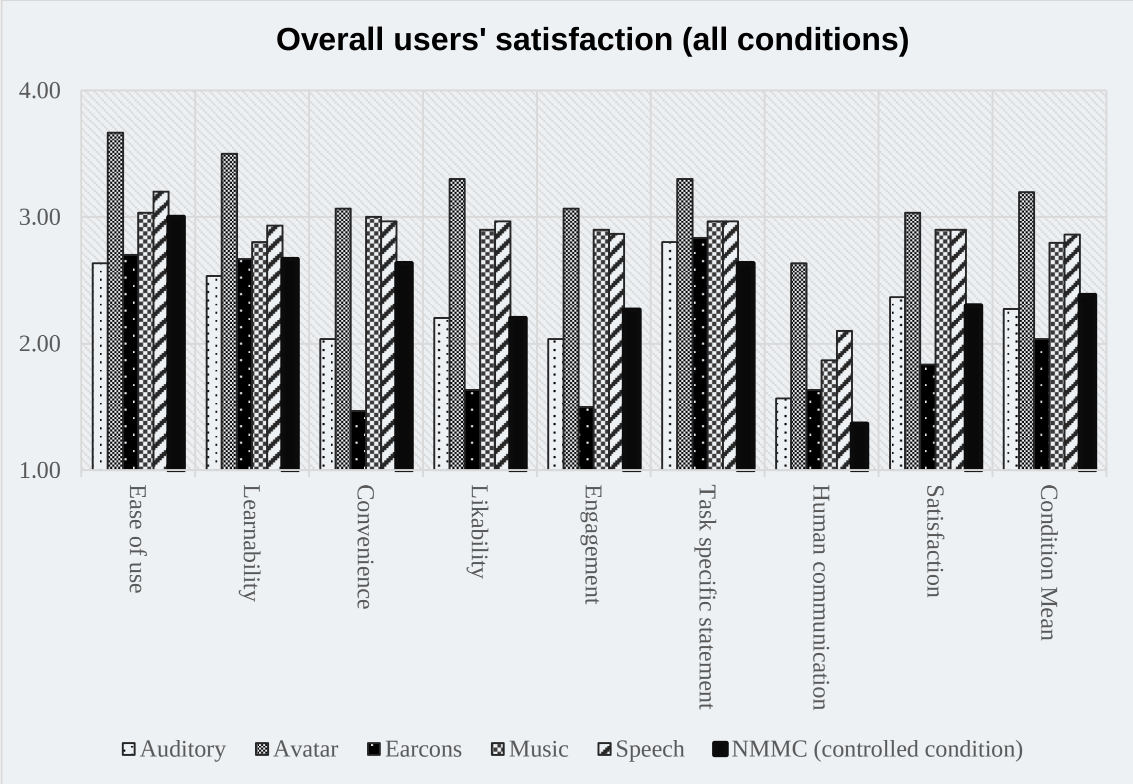


Figure 4-30: Overall users' satisfaction results.

4.6.3.7 ***Statistical evaluation***

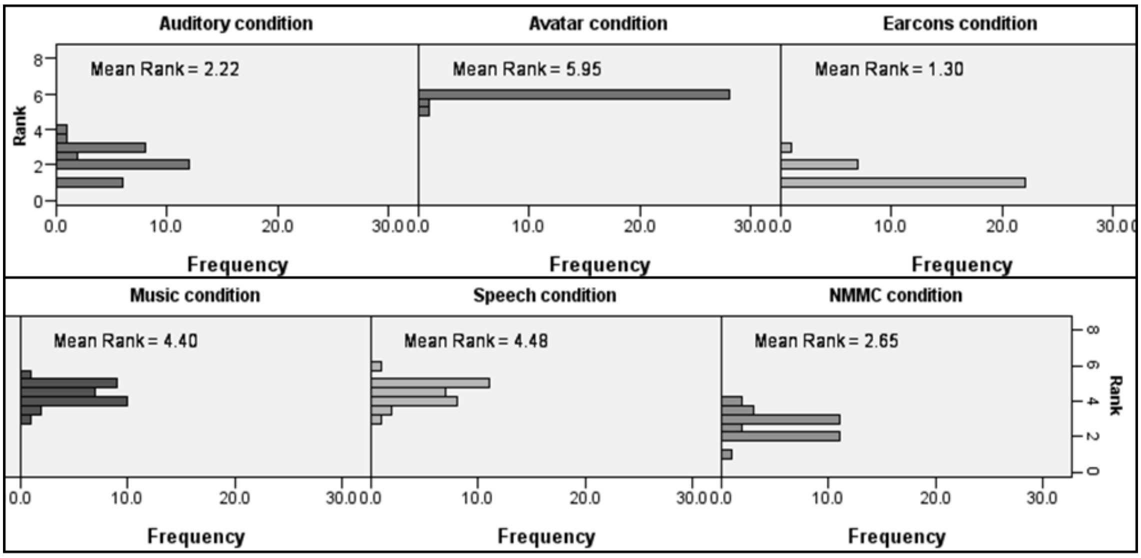
When performing statistical analysis it is crucial to select an appropriate test according to the distribution of the data to be tested, therefore a normality test was performed using "Shapiro-Wilk" to check whether the satisfaction data was normally distributed. The result is listed in Table 4-8.

Tests of normality						
Condition	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.145	30	.110	.951	30	.183
Avatar condition	.154	30	.069	.925	30	.037
Earcons condition	.175	30	.020	.933	30	.059
Music condition	.236	30	.000	.899	30	.008
Speech condition	.160	30	.049	.947	30	.144
NMMC condition	.238	30	.000	.858	30	.001
a. Lilliefors Significance Correction						

Table 4-8: Normality test (users satisfaction data) results.

The test results confirm that the data is not normally distributed and therefore a non-parametric statistical test will be used, taking into account the study design and the number of samples. The related-samples Friedman's two-way Analysis of Variance (ANOVA) by rank is the recommended test to be used. The Friedman's ANOVA test result confirms that the differences between the tested conditions in terms of user satisfaction were statistically significant ($\chi2$ (5) = 131.226, p < .0005.). For detailed test results, see Figure 4-31. Unfortunately, Friedman's ANOVA test did not show where the differences are among the tested conditions; consequently, a *post-hoc* test was needed, thus a pairwise comparison with Bonferroni adjustment for multiple comparisons was performed.

Effectiveness related samples Friedman's ANOVA test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory condition, Avatar condition, Earcons condition, Music condition, Speech condition and NMMC condition are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

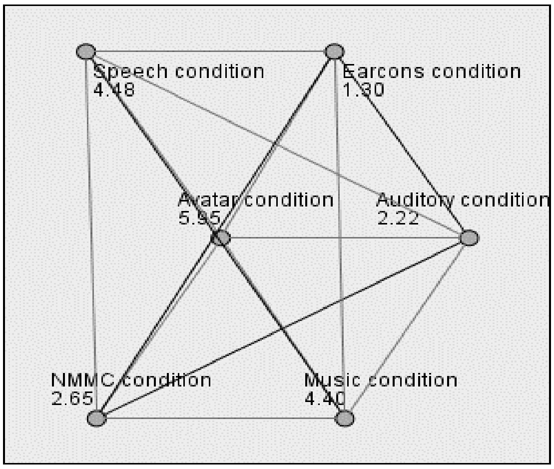
Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	131.226
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 4-31**:** Users satisfaction related-samples Friedman's ANOA test.

The full test results are presented in Figure 4-32 and the significant differences are denoted in gray. It is possible to divide the tested conditions into three levels. The first level includes the avatar condition with a higher number of significant differences than the remaining (including the controlled) conditions.

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Earcons condition-Auditory condition	.917	.483	1.898	.058	.866
Earcons condition-NMMC condition	-1.350	.483	-2.795	.005	.078
Earcons condition-Music condition	-3.100	.483	-6.418	.000	.000
Earcons condition-Speech condition	-3.183	.483	-6.590	.000	.000
Earcons condition-Avatar condition	4.650	.483	9.626	.000	.000
Auditory condition-NMMC condition	-.433	.483	-.897	.370	1.000
Auditory condition-Music condition	-2.183	.483	-4.520	.000	.000
Auditory condition-Speech condition	-2.267	.483	-4.692	.000	.000
Auditory condition-Avatar condition	-3.733	.483	-7.729	.000	.000
NMMC condition-Music condition	1.750	.483	3.623	.000	.004
NMMC condition-Speech condition	1.833	.483	3.795	.000	.002
NMMC condition-Avatar condition	3.300	.483	6.832	.000	.000
Music condition-Speech condition	-.083	.483	-.173	.863	1.000
Music condition-Avatar condition	1.550	.483	3.209	.001	.020
Speech condition-Avatar condition	1.467	.483	3.036	.002	.036

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 4-32: Pairwise comparison.

The second level includes the speech as well as the music conditions, both of which were significantly different from the other conditions (including the controlled ones); however, the difference between them did not reach a statistically significant level. The last level includes the earcons condition as well as the auditory condition and the control condition; these were significantly different from those in the above-mentioned levels and the differences among them did not reach a statistical significance. For a better understanding of the pairwise comparison, test results Figure 4-33 shows the conditions ranked according to users' satisfaction means and medians in ascending order.

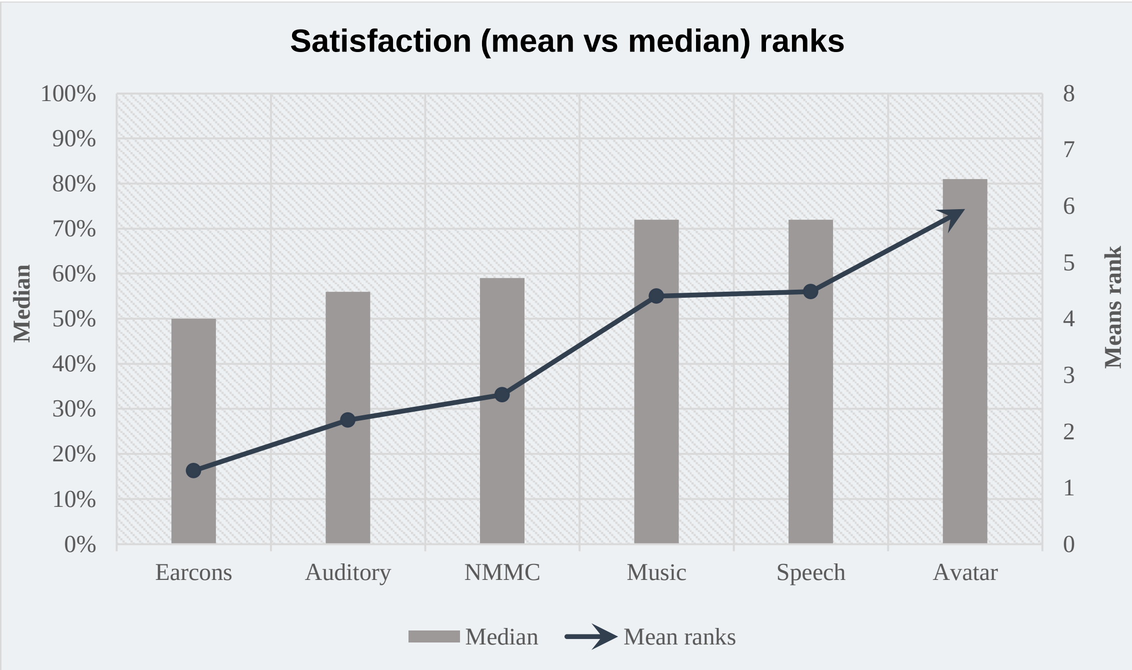


Figure 4-33: Pairwise comparison (means and medians).

4.6.4 Discussion

The assessment results reveal the impact that the employed communication metaphor had on usability of the tested conditions compared to the control condition. The results clearly show that using a particular communication metaphor could improve or deteriorate the overall system usability or some of its attributes.

4.6.4.1 *Efficiency*

In terms of tasks' efficiency, the results presented in Figure 4-10 indicate that using avatar, speech and music improved the overall system efficiency in performing the presented tasks. However, there were no significant variances due to the inclusion of auditory icons and the earcons. Concerning efficiency in relation to task type, the results presented in Figure 4-8 as well as Figure 4-9 indicate that , the inclusion of the communication metaphors (avatar, speech and music) led to improvements in performing recall and recognition tasks. Furthermore, task performance efficiency increased with the increase in task cognition load (this is shown in Figure 4-10, as the decrease in differences between recall and recognition task efficiency). From the results it seems that ,the addition of the communication metaphors in the avatar experimental conditions aided the users to focus much better on the demonstrated information over the auditory channel while concurrently utilizing the visual channel to comprehend them [206]. The obtained outcomes proposed that using the facially expressive speaking avatar, classic music, recorded speech and auditory icons could be more efficient than using only the text with graphic metaphors (NMMC) “visual only and non-verbal” in the presentation of the learning topic in this experiment. There was only one exception to that in the earcons experimental condition, where it seemed that the inclusion of these communication metaphors was less efficient than the control condition; furthermore, the results showed that using earcons alone could have a negative effect on the system efficiency. This is in contrast to the result obtained in the first experimental phase (Chapter 3) where the earcons communication metaphor was utilized in a multimodal setup in conjunction with other communication metaphors. It is also worth mentioning here that, the observed improvements were not equal across all the experimental conditions; thus, partly accepting what has been hypothesized in H1 and rejecting what was predict in H2.

4.6.4.2 *Effectiveness*

In the evaluation of including the communication metaphors on effectiveness, the results (Figures 4-12 and 4-15) show that the avatar, speech, music and auditory icons enhanced the overall task performance efficiency. Concerning effectiveness in relation to task type, the results shown in Figures 4-13 and 4-14 indicate that improvement in effectiveness was prominent in recall as well as recognition tasks and it is noted that with the inclusion of avatar communication metaphor the effectiveness increased with increased cognition loads in recall tasks. The experimental conditions effectiveness evaluation results have shown that utilizing additional communication metaphors to those of NMMC could assist users to perform better going through tasks with aggregated complexity, where extra information is conveyed and a lesser amount of cognitive-resources come to be obtainable while processing them [27]. The results obtained from experiment suggests that using multimodal communication metaphors possibly will help in prolonging the capacity of working-memory to empower the user's ability to process both auditory and visual information [207]. The results shown that the obtained improvements in terms of system effectiveness were not the same across all the experimental conditions, which indicate that communication metaphors were not equally effective therefore H3 is rejected. Furthermore, within the earcons experimental condition there were no noticeable improvements concerning system effectiveness in comparison to the control condition, therefore H4 is partially rejected. Concerning task type and users' performance, the observed improvements varied among tasks in connection to their type and complexity levels and the utilized communication metaphor. In recognition tasks, there were noticeable improvements in users' scores within the avatar, speech, classic music and recorded speech in comparison to the control condition, the highest improvement was attained within the avatar condition followed by the speech and classic music conditions while the least improvement was in the auditory icons condition. However, for the earcons condition performance in the recognition category it was disappointing and less than that for the control condition; it is worth mentioning here that improvements decreased as the complexity increased across all the exponential conditions. With regards to the recall tasks where the cognition load was higher, the experimental conditions with shown were persistent in their performance and coped well with the increased complexity levels of the presented tasks especially in the avatar experimental condition. it seems that the using the multimodal communication metaphors have the increased users ability to retain the presented the presented information longer with the exception to the earcons metaphor were users seemed to be anxious while they were in use; this indicates that users comfort can have an impact on their performances and their ability to retain or acquire information. Consequently, H5 is rejected.

4.6.4.3 *Learning performance*

For the evaluation of interaction metaphors' influence on users' learning performance, Figure 4-19 depicting users' learning time as well as their achieved scores indicates that there were improvements with the use of avatar, speech, and music communication metaphors. However, there were no visible enhancements while using auditory icons and a decrease in learning performance with the use of earcons. In addition, the correlation results in Figure 4-20 clearly confirm that the observed enhancements were due to the inclusion of the communication metaphors and not to the increase in the learning time in comparison with the control condition (NMMC).

The obtained improvements in the learning performance can be attributed to the multimedia-principle [208] [209]; where encompassing additional human senses to the visual channel in the communication route possibly will help in widening the working memory's capacity and reducing the learning time; subsequently, the users' ability to observe and apprehend the communicated information can be enhanced. With reference to the results obtained in this evaluation category, what was hypothesised in H6 is rejected.

4.6.4.4 *User satisfaction*

Concerning users' satisfaction, the inclusion of the communication metaphors (avatar, speech and music) was very positive (see Figures 4-21 and 4-30) It is also noted that regarding the avatar communication metaphor, users' missing of human communication was less severe than in other metaphors followed by the speech and music metaphors respectively. On the other hand, with other communication metaphors (earcons and auditory) users were less satisfied in comparison with the control condition. The biggest downfall for these metaphors was noted in the areas of "convenience of use" (Figure 4-24), "likeability" (Figure 4-25) and "engagement" (Figure 4-26) and "human communications" (Figure 4-27).

Ease of use: Users rated the experimented conditions concerning their observed ease of use and the results are depicted in Figure 4-22. All the experimented conditions were praised in this category, where the experimental conditions can be congregated into two groups. The first group contains two conditions (earcons, auditory icons) where experimental conditions' "ease of use" was rated as less than the control condition and the second group of conditions (avatar, recorded speech and classic music) were rated higher. There was 91% agreement among users that the avatar condition was the easiest to use, followed by the speech condition with 80% rating and the classic music condition with 75.8% in that order. The lowermost ranking was given to auditory with 65.8% trailed by the earcons conditions with 67.5%; the evaluation results obtained from the users' point of view clearly indicate that using certain communication metaphors could improve or worsen users' satisfaction with the system's perceived ease of use.

Learnability: Users rated how easy it was to learn to use each experimental condition. The results presented in Figure 4-23 show the avatar condition as the easiest to learn with 78.5% agreement, followed by the speech condition with 73.30% and the classical music condition with 70% agreement. On the other hand there was no difference observed with the earcons condition in comparison with the control condition where both scored 66.75%, while the most difficult to learn condition proved to be the auditory one with 63.25% rating. The results clearly show that using certain communication metaphors could help to improve system learnability.

Convenience of use: Referring to Figure 4-24, which shows how comfortable users were using each of the experimented conditions, it can be seen that some conditions performed better than the control condition while others performed worse. The lowest ranking was for the earcons condition at 36.75%, followed by the auditory condition, which was rated 50.80%; both conditions were shown to be less convenient to use than the control condition, which was much higher at 65.75%. Users elected the avatar condition as the most convenient condition to use with 76.80% rating, followed by the classical music condition that received 75% rating and speech condition, which scored 74.23% in this category. The results obtained in this satisfaction class clearly indicated that utilizing particular communication metaphors could make the system a convenient smooth experience or an annoying and uncomfortable one.

Likability: Figure 4-25 shows users' interests and to what extent they liked each of the inspected conditions. They rated the avatar condition as the most preferred condition (rated=82.5%), followed by speech (rated = 74.25%) and classical music (rated = 72.5%). The auditory icons and the control condition were equally rated (rated = 55%) and the lowermost rating was for the earcons condition (rated = 41.75%). The assessments in this satisfaction category clearly designated that the inclusion of certain communication metaphors could help improve users' fondness of the system.

Engagement: Figure 4-26 presents how users perceived their engagement with experimented conditions. They rated the avatar condition (rated = 76.75%) as the most engaging condition, followed by the speech and music conditions as they were relatively equally engaging and performed much better than the control condition.

The next rank was for the control condition (rated = 56.75%), followed by the auditory condition (rated = 56.75%) and then the least engaging condition (rated=37.50%). The outcomes of this satisfaction category assessment clearly show that using some communication metaphors could increase or decrease users' interest in the system presented to them, which could greatly affect their performance as well as their long-term adoption of such a system

Human-to-human communications: Users rated to what extent they felt isolated and missed human-to-human social skills while experimenting with presented conditions and the results were presented in Figure 4-27. The condition where users had the worst experience was with the control condition (rated = 34.25%); all the experimental conditions outperformed the control condition in this satisfaction category. In comparison to control condition, auditory condition was rated 5% higher, then the earcons condition 6.5% higher and the music conditions 12.5% higher. Users' best rating was given for the avatar condition, which scored 31.5% higher than the control condition followed by the speech conditions, which scored 18.25% higher. The results clearly indicated that using certain communication metaphors could be used to some extent to compensate for the absence of human-to-human social skills while interacting with the system.

Condition specific satisfaction statement: Figure 4-29 presents the users satisfaction with specifically formulated satisfaction statements related to the type of communication metaphor utilized in each experimental condition. Users were relatively high in agreement with statements regarding all conditions; concerning the control condition users (with 65.75% agreement) thought that without sound they were able to concentrate better. In the auditory icons condition statement (with 70% agreement) users thought that the addition of auditory icons made the lesson more realistic. With regards to the earcons condition statement, users thought (with 70.75% agreement) that the earcons alerted them to the important points. In the music condition statement, users assumed (with 74.25% agreement) that, the music helped them to relax and concentrate. Concerning the speech condition statement, users thought with same level of agreement that without having to read texts, they were able to concentrate better. Concerning the avatar condition special satisfaction statement, users thought with (82.50% agreement) that the virtual lecturer helped them to understand the presented learning topic. The obtained results clearly indicate that each communication metaphor has a unique feature that could be used to improve the system usability, even for those metaphors that did not perform well in other satisfaction categories, when they were used in the first experimental phase (3.0) in conjunction with other communication metaphors they performed well. The results indicate that certain interaction metaphors could perform better when they are utilized in a multimodal communication metaphor scenario.

Overall users' satisfaction: Users were requested to rate their approval and satisfaction with each experimented condition and the results are depicted in Figure 4-29. Users voiced the highest satisfaction for the avatar condition (75.75%) followed by the speech and music conditions, which were equally satisfying (72.20%). The auditory icons were next in the rank and 1.8% higher than control condition (57.50%). Overall, the earcons condition was the least satisfying condition and was rated 11.80% lower than the control condition. The results clearly indicate that the use of certain communication metaphors could be utilized to improve users' satisfaction with the system, which would have a positive impact on their performance as, was clearly shown in the results seen in sections 4.6.3.1, 4.6.3.3 and 4.6.3.5. Based on the aforementioned results what was hypothesised in H7 is rejected.

4.1 SUMMARY

A usability evaluation study was conducted for an e-learning platform with five different experimental conditions (each condition dominated by a single interaction metaphor) in addition to one control condition. The aim of the study was to investigate the impact of each metaphor on the system's overall usability. Usability was evaluated in four areas (efficiency, effectiveness, learning performance and users' satisfaction). The evaluation method involved calculating users' learning times, task performance times, task performance scores and subjective responses to the satisfaction statements. The best usability improvements were perceived when using the avatar interaction metaphor, followed by the recorded speech, then the music metaphor and then the auditory icons. The least usability was witnessed while using the earcons metaphor. Furthermore, the results show that certain communication metaphors, such as the facially expressive speaking avatar, recorded speech and classical music are of great help to system usability and it would be useful to include them more often. Nonetheless, other metaphors (auditory icons and earcons) should be used with caution due to their acoustic strength, which might hinder system usability and result in a negative impact on users' learning performance. The results clearly show that system usability and users learning performance could significantly improve with the inclusion of additional communication metaphors, where more human senses are employed to communicate with the user. Furthermore, with reference to the results obtained while using earcons as the dominating communication metaphor in comparison with the results obtained in the first experimental phase, the results show that employing a multimodal metaphoric approach could be more beneficial and reduce the side effects of over using one or more of the interaction metaphors. The results in other evaluation areas presented in Chapter 5 and Chapter 6 will shed more light on how the utilized communication metaphors influenced the users' experience and affective states. The overall usability results are summarized in the usability traffic light (see Figure 4-34).

	Usability trafic light				
Condition	Efficiency	Effectiveness	Satisfaction	Learning Performance	Usability (Overall)
Auditory	!	!	✗	!	✗
Avatar	✓	✓	✓	✓	✓
Earcons	✗	✗	✗	✗	✗
Music	✓	✓	✓	✓	✓
Speech	✓	✓	✓	✓	✓
NMMC	!	!	!	!	!

Figure 4-34: Usability evaluation traffic light.

CHAPTER 5

5 USER EXPERIENCE EVALUATION (EXPERIMENTAL PHASE II)

User experience evaluation aims at assessing the experienced conditions in terms of usefulness, user-friendliness, stimulation and attractiveness. User experience is evaluated within three dimensions, pragmatic quality, hedonic quality (identification and stimulation) and attractiveness.

User experience evaluation differ from usability evaluation (Chapter 4) because it goes beyond the traditional concept of user satisfaction in usability evaluation to a more holistic overview exploring the aspects of the experimented conditions in terms of their attractiveness ,pragmatic and hedonic user experience.

5.1 AIMS and OBJECTIVES

The principal aims of the experiment within the user experience dimension are to:

- Investigate the impact of each investigated communication metaphor on the overall user experience.
- Test five experimental conditions where communication metaphors are individually employed.
- Assess users' pragmatic experience with each experimented condition.
- Assess users' hedonic experience within each experimented condition.
- Assess the experimented conditions' attractiveness.
- Test whether there are any significant differences between the experimented conditions within the users' in terms of :
    - Pragmatic user experience.
    - Hedonic user experiences.
    - Attractiveness.
- Identify which condition achieved the highest performance in terms of user experiences.

Accomplishing the aims of this experiment requires that, the following objectives are met:

- Formulate user experiences experimental hypotheses.
- Test the following conditions:

- Auditory condition.
    - Music condition.
    - Earcons condition.
    - Avatar condition.
    - Speech condition.
    - Control condition.
- Measure the user experience of the tested conditions in terms of their perceived:
    - Pragmatic quality.
    - Hedonic quality-identification.
    - Hedonic quality-stimulation.
    - Attractiveness.

5.2 HYPOTHESES

It is anticipated that the utilized interaction metaphors will have an equal positive influence on users' experience and the following hypotheses have been formulated accordingly:

H1: The experimental conditions will receive equal preferences and rating in their provided pragmatic user experience measurements.

H2: The experimental conditions will receive equal preferences, users identification and rating in terms of their provided hedonic user experience.

H3: The experimental conditions will receive the same preferences as well as users' stimulation and rating within their provided hedonic user experience.

H4: The experimental conditions will be equally appealing and attractive in their provided user experience measurements.

5.3 METHOD OF INVESTIGATION

The investigation method adapted in this research to assess users experience was subjective in nature utilizing AttrakDiff ™ instrument [15, 210]. Assisted by pairs of opposite-words, users indicated their experience with the tested conditions within three-targeted areas, which include pragmatic as well as hedonic qualities and attractiveness. AttrakDiff reports the mean value of the hedonic and pragmatic quality results of the experimented conditions based on users' estimation of hedonic qualities, which are denoted along the vertical axis, whereas the horizontal axis denotes pragmatic quality estimated value.

According to the dimensions values a particular condition is placed in one of the “character-regions”.

AttrakDiff confidence rectangle: the larger the rectangle is the less sure one can be as to which region it belongs. A small confidence rectangle is an advantage because it means that the investigation results are more reliable and less coincidental. The confidence rectangle shows if the users are at one in their evaluation of the platform).

5.4 EVALUATION RESULTS

5.4.1 Auditory Icons Condition User Experience

The auditory icons condition was rated, as neutral”, pragmatic quality is clearly the classification. Users were assisted by the condition; on the other hand, the rating of pragmatic quality just meets the ordinary levels (see Figure 5-1).

Figure 5-1: PQ and HQ medium values (auditory icons condition).

Thus, there is an opportunity for enhancement within the pragmatic dimension. Concerning the hedonic quality, the condition is positively classified. The users were stimulated by it; though, the scored hedonic value is only reaches the average values. Consequently, there is a chance for expansion within the dimension of hedonic qualities.

Having a size and direction of the confidence rectangle indicates that, users were at large agreement in their rating of the condition's pragmatic quality than their rating of hedonic qualities.

5.4.1.1 *Auditory icons mean values*

The mean values of the AttrakDiff dimensions for evaluating auditory icons are depicted in Figure 5-2. Here, the stimulation and identity aspects of hedonic quality are distinguished between. Additionally, the rating of the condition's attractiveness is presented.

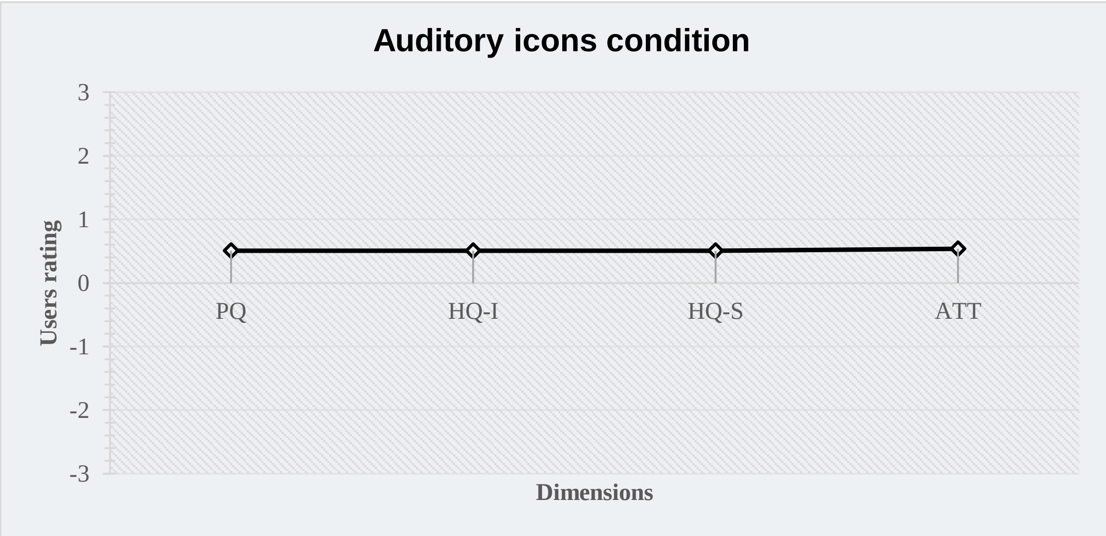


Figure 5-2: Mean values (auditory icons condition).

Hedonic quality – identity: This condition is positioned within the average region; users identified with condition which in this category meets ordinary standards. To strive and encourage the user to continue using it an enhancement is obligatory.

Hedonic quality – stimulation: The auditory icons condition is situated within the average region, it satisfies ordinary standards; nevertheless to motivate, captivate and stimulate users even more intensely, there is a need for further improvement.

Attractiveness: The attractiveness score is positioned within the average region. The overall assessment of the condition within this category is neutral.

5.4.1.2 *Auditory Icons (description of word-pairs)*

The condition users rating of the word-pairs are presented in Figure 5-3, the figure clearly shows the areas which are well addressed or in need of an urgent attention.

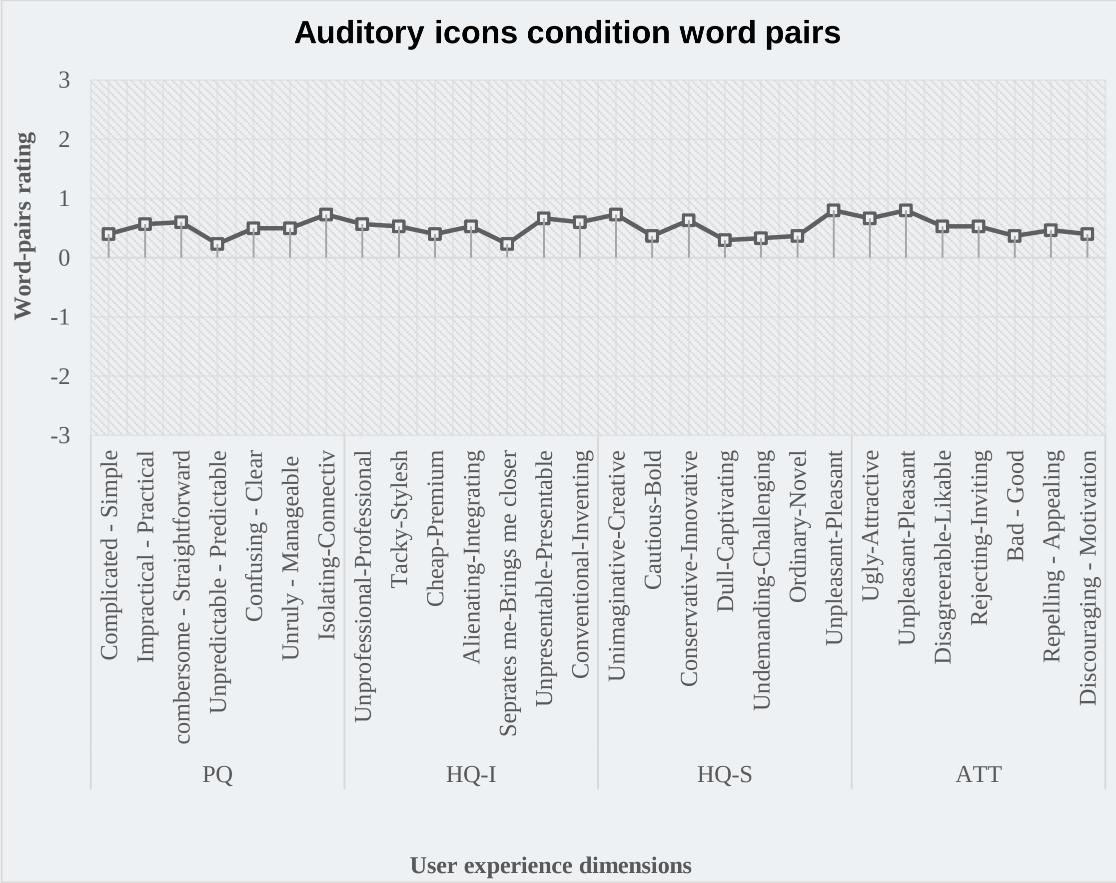


Figure 5-3: Word-pairs (auditory icons condition).

With regards to auditory icons' PQ, to some extent users connected well with this condition, their rating was slightly above average, they thought it was slightly simple, practical, clear and manageable. With regards HQ-I users from their experience thought thee condition was professional, to some degree stylish, premium and presentable. Concerning HQ-S, users were marginally stimulated by the condition; they considered it to be creative, bold, innovative and pleasant. In terms of condition attractiveness, it was rated slightly above the average, considered to be pleasant, likable, useful and not discouraging.

5.4.2 Avatar condition User Experience

The avatar condition was rated as "rather desired"; users were assisted by the condition and pragmatic quality is clearly the classification; the value of pragmatic quality was higher than the average value. However, still there is an area for improvement in terms of the condition pragmatic quality (See Figure 5-4 for the mean values of the hedonic and pragmatic qualities).

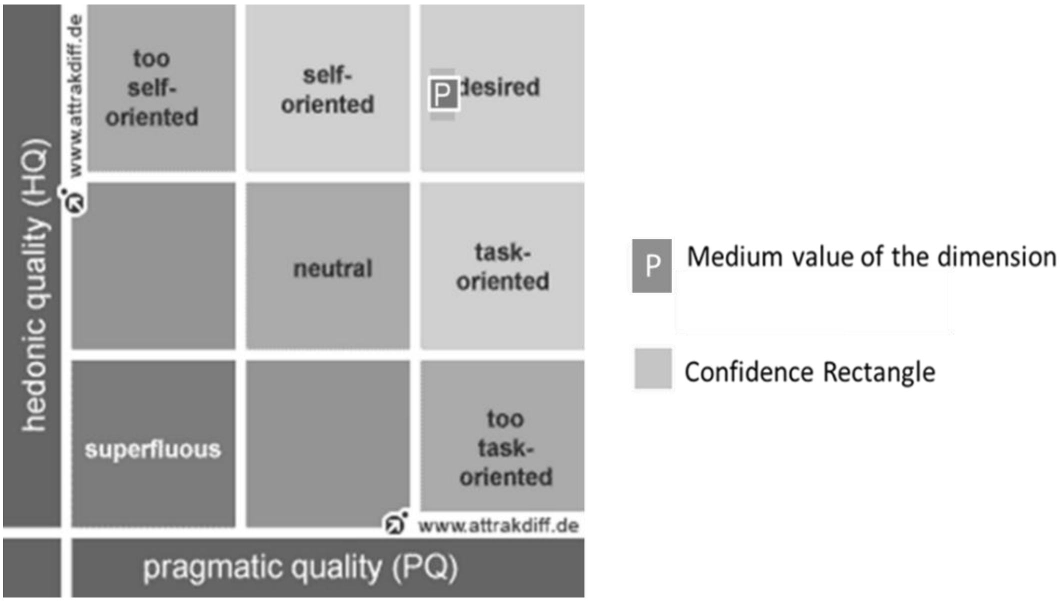


Figure 5-4: PQ and HQ medium values (avatar condition).

With regards of hedonic quality, the classification clearly applies to the condition. Users deemed it as highly hedonic. Users identified themselves very well with the condition and were highly encouraged and stimulated by it.

The small size of the confidence rectangle clearly indicates that users were at one in their rating of the condition.

5.4.2.1 *AVATAR Mean Values*

AttrakDiff dimensions mean values for the avatar condition are shown in Figure 5-5, which distinguishes between the hedonic quality aspects of stimulation and identity. Moreover, the condition attractiveness score of is presented.

Hedonic quality – identity: this condition is placed in the above-mean region. It offers users with identification and consequently binds them to the condition. Overall, within the identity aspects, this condition is deemed as optimal.

Hedonic quality – stimulation: this condition is placed in way higher position than average. It motivates users, awakes interest and inspires them. Overall, with regards to this category, this condition is classified as optimal.

Attractiveness: This condition’s attractiveness score is placed in apposition higher the average region. The total impression of this condition is highly attractive.

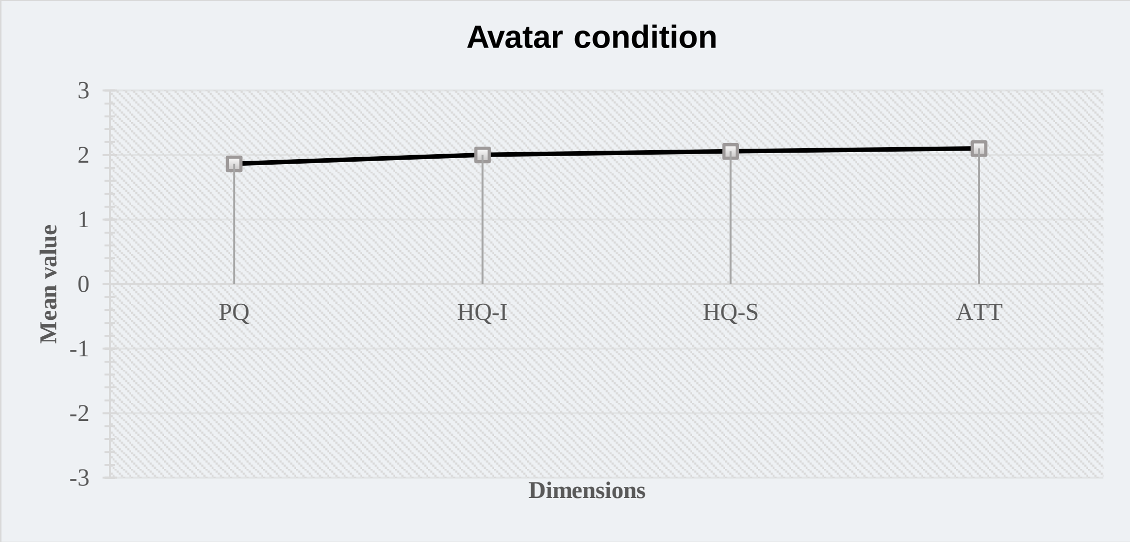


Figure 5-5: AttrakDiff dimensions' mean values (avatar condition).

5.4.2.2 *Description of Word-Pairs*

Figure 5-6, clearly shows the areas which are well addressed or in need of an urgent consideration.

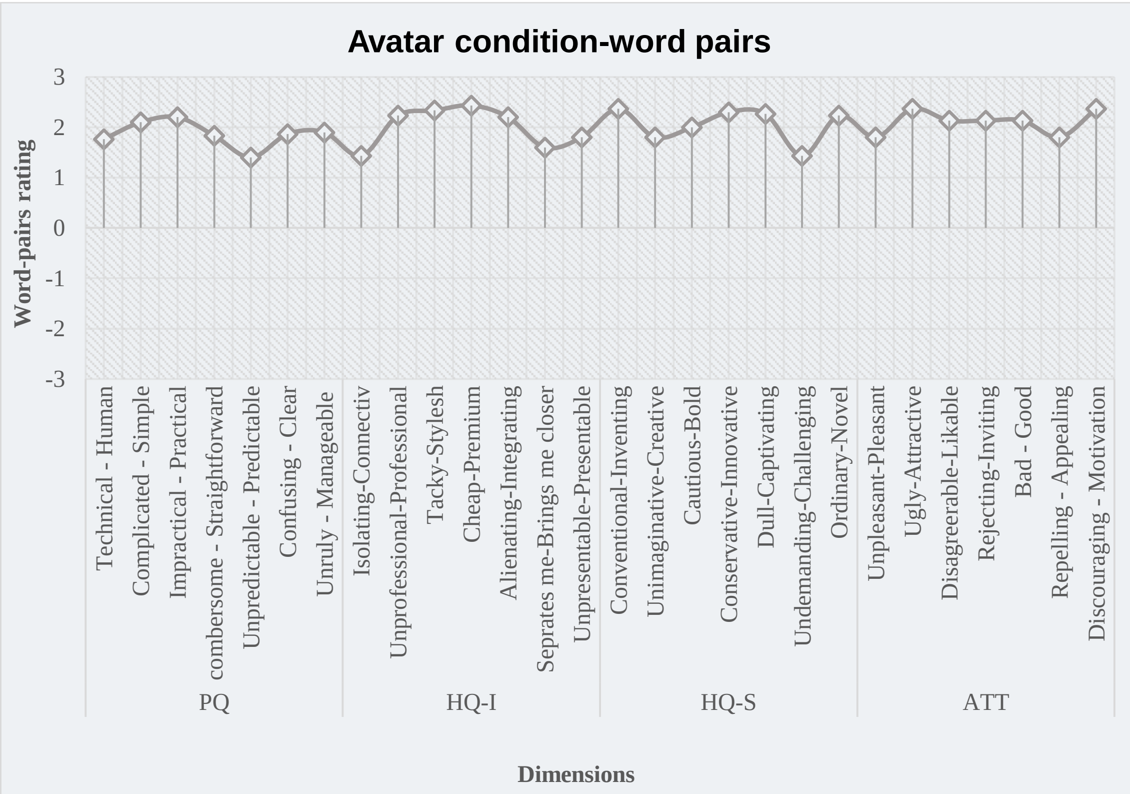


Figure 5-6: AttrakDiff (word-pairs) for avatar condition.

Concerning the avatar condition PQ, users connected very well with this condition as they thought it was more human than technical, their rating was high and they thought it was simple, useful, vibrant and manageable. With respect to the condition's HQ-I, users rated their experience as with high identification, they thought it was professional, very stylish and exceptionally presentable. Concerning HQ-S, users were highly stimulated by the condition; they thought of it as innovative, bold, captivating, novel and pleasant. In terms of condition attractiveness, it was rated as highly attractive, pleasing, friendly, beneficial and encouraging.

5.4.3 Earcons Condition User Experience

This condition was rated as "neutral". The classification pragmatic quality is the classification. The users were slightly aided by the condition; conversely, the value of pragmatic quality was just below average. Subsequently, at hand there is definite opportunity for enhancements, (see Figure 5-7). Concerning the condition hedonic quality, it did not fit the classification. Users were stirred by it; however, the hedonic value was below average. Therefore, within this category there is a large room for improvements. The confidence rectangle is large and crosses to another character region; this indicates that, users rating for the condition within this category widely varied, user were not in agreement over their rating of the conditions' hedonic quality.

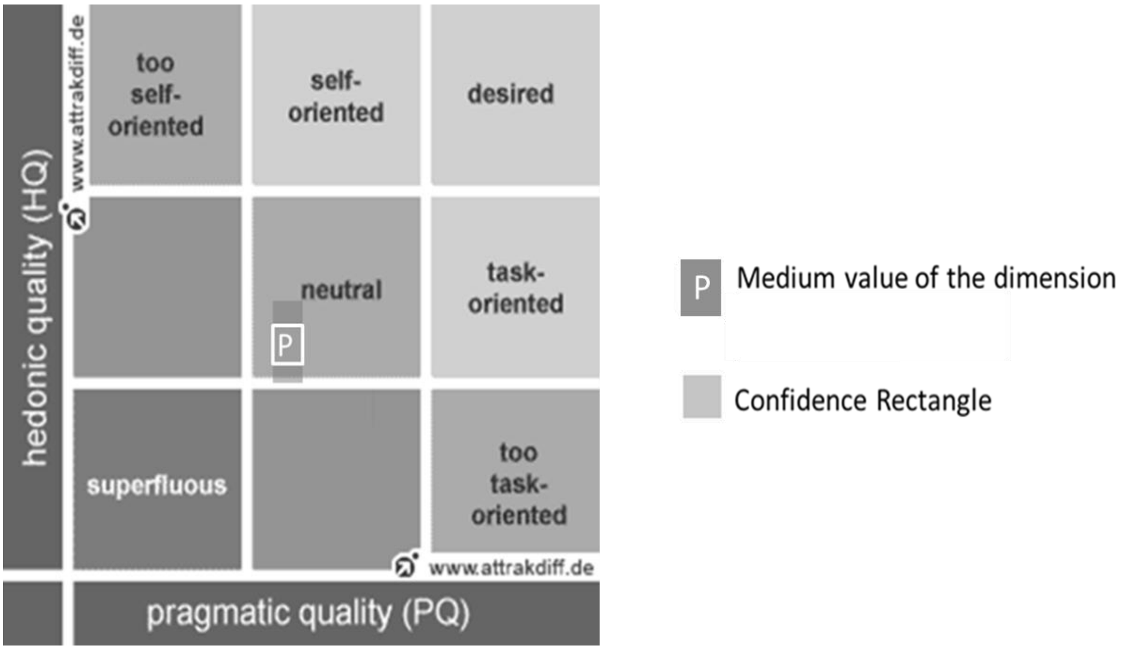


Figure 5-7: PQ and HQ medium values (earcons condition).

5.4.3.1 *Mean values of users experience*

The mean scores of the AttrakDiff measurements for the earcons condition are shown in Figure 5-8, in which, the stimulation and identity aspects of hedonic qualities are distinguished. Furthermore, the rating of attractiveness is presented.

Hedonic quality – identity: This condition is positioned below the average region and it is in need for urgent attention. To bind the user to this condition, improvement is necessary.

Hedonic quality – stimulation: This condition is placed within the average region and it just met ordinary standards. To motivate, absorb and stimulate users further improvement is necessary.

Attractiveness: This condition's attractiveness value is located way below the average. The whole impression of the condition is unattractive.

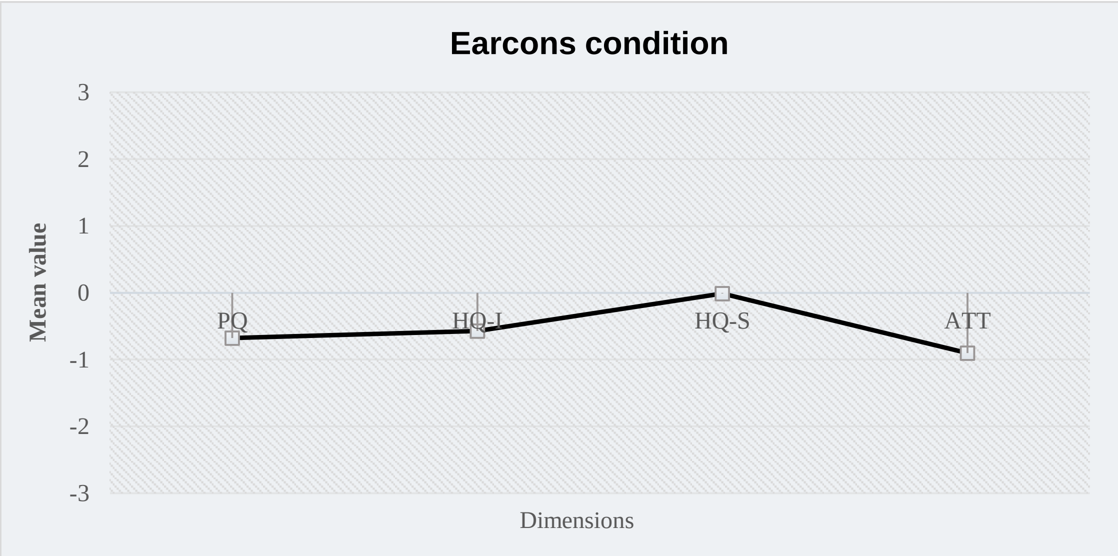


Figure 5-8: AttrakDiff dimensions' mean values (earcons condition).

5.4.3.2 *Earcons Description of Word-Pairs*

The mean values of the word-pairs are presented here. These show which features are critical or well-resolved (see Figure 5-9). Concerning earcons' condition PQ, users rating was below average, to some extent users did not connect well with this condition, they thought it was technical, cumbersome to use, impractical and unruly. Regarding HQ-I, users thought the condition was isolating, to some degree unprofessional, tacky, cheap and unpresentable. Concerning HQ-S, users were not inspired by the condition; they considered it ordinary, slightly dull and cautious. In terms of condition attractiveness, it was rated below the average, it was considered unpleasant, ugly, rejecting and discouraging. Accordingly, the condition is in an urgent need for revision.

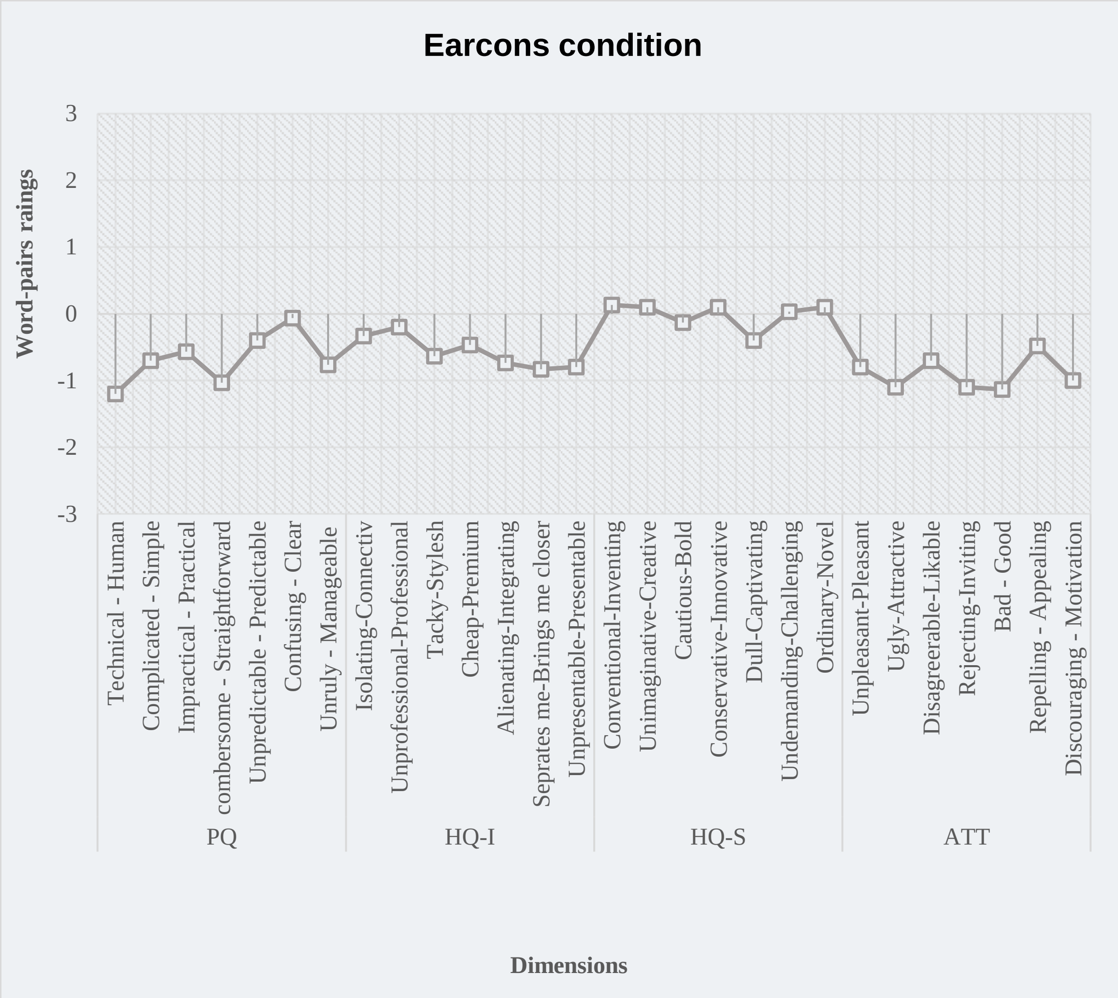


Figure 5-9: AttrakDiff word-pairs (earcons condition).

5.4.4 Classic Music Condition User Experience

The music condition was deemed as confidently “neutral”. The classification of pragmatic quality is evident. The users are helped by the condition; however the value of pragmatic quality was moderate; accordingly, there is room for improvement within this category (see Figure 5-10). In terms of hedonic quality, the character classification applies positively. The users were stimulated by this condition; however, the hedonic value is moderate. There is possibility for expansion within this category. The confidence rectangle is small which indicates that, users were at one in their evaluation of the condition.

5.4.4.1 *Mean values of users experience*

The mean values of the AttrakDiff dimensions for the classic music condition are illustrated together in Figure 5-11, in which the stimulation and identity aspects of hedonic quality are distinguished. Furthermore, within users ranking of the condition, attractiveness was found to be the highest and most appealing dimension.

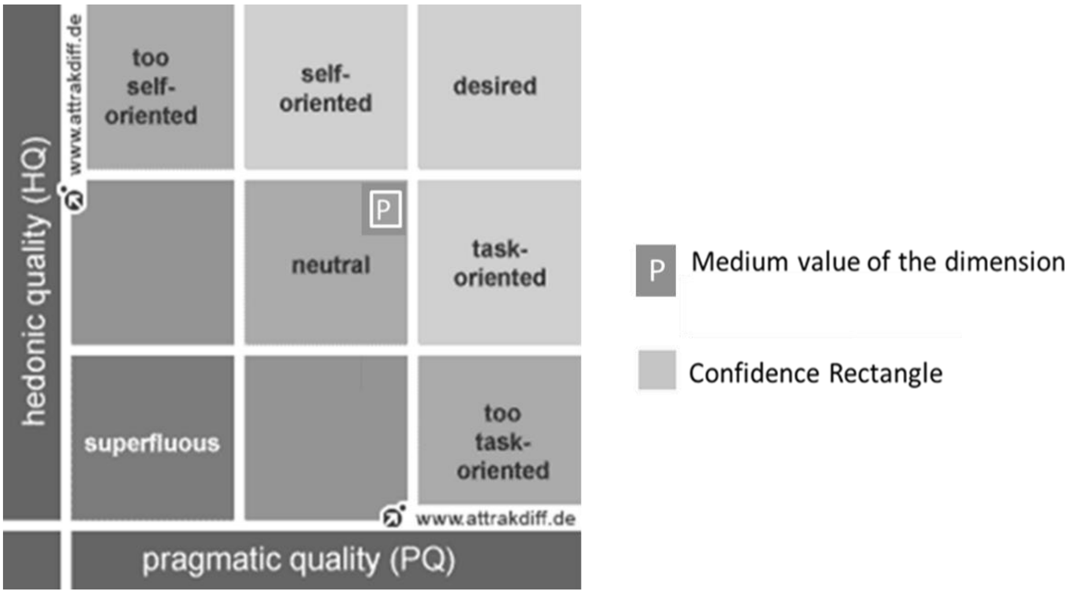


Figure 5-10: PQ and HQ medium values (music condition).

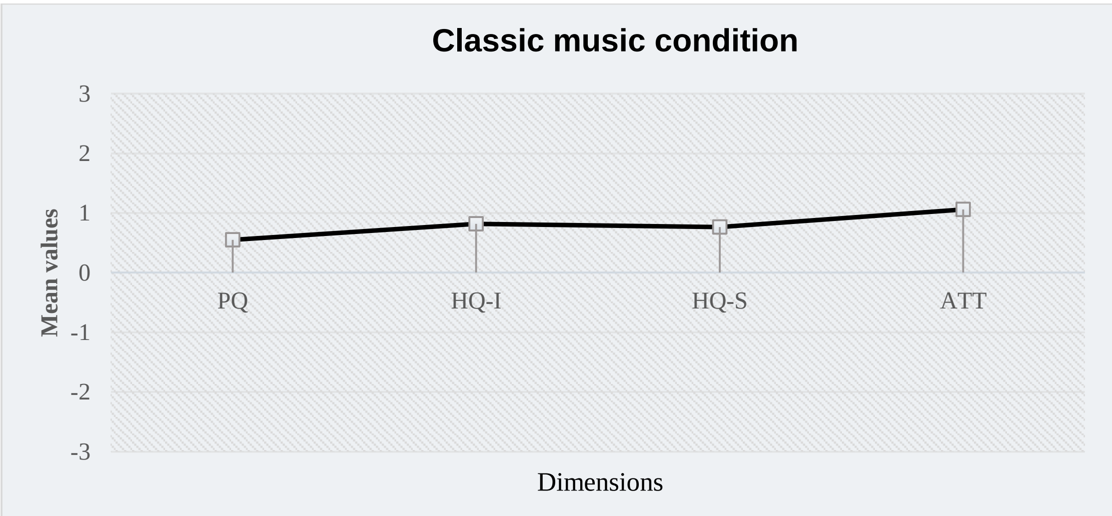


Figure 5-11: AttrakDiff dimensions’ mean values (music condition).

Hedonic quality – identity: This condition is placed in the above average region and users identified themselves well with the condition. The condition meets ordinary standards; to bind the user more strongly to the condition further improvement is necessary.

Hedonic quality – stimulation: This condition is placed in the mean region. It meets ordinary standards; to motivate attraction and stimulate users even more intensely further improvement is necessary.

Attractiveness: This condition's attractiveness score is placed in the above average region; the total impression of the music condition is moderately attractive.

5.4.4.2 *Classic Music Description of Word-Pairs*

AttrakDiff word-pairs are presented in Figure 5-12 , this show that the areas, which were mainly critical or overall, were well resolved.

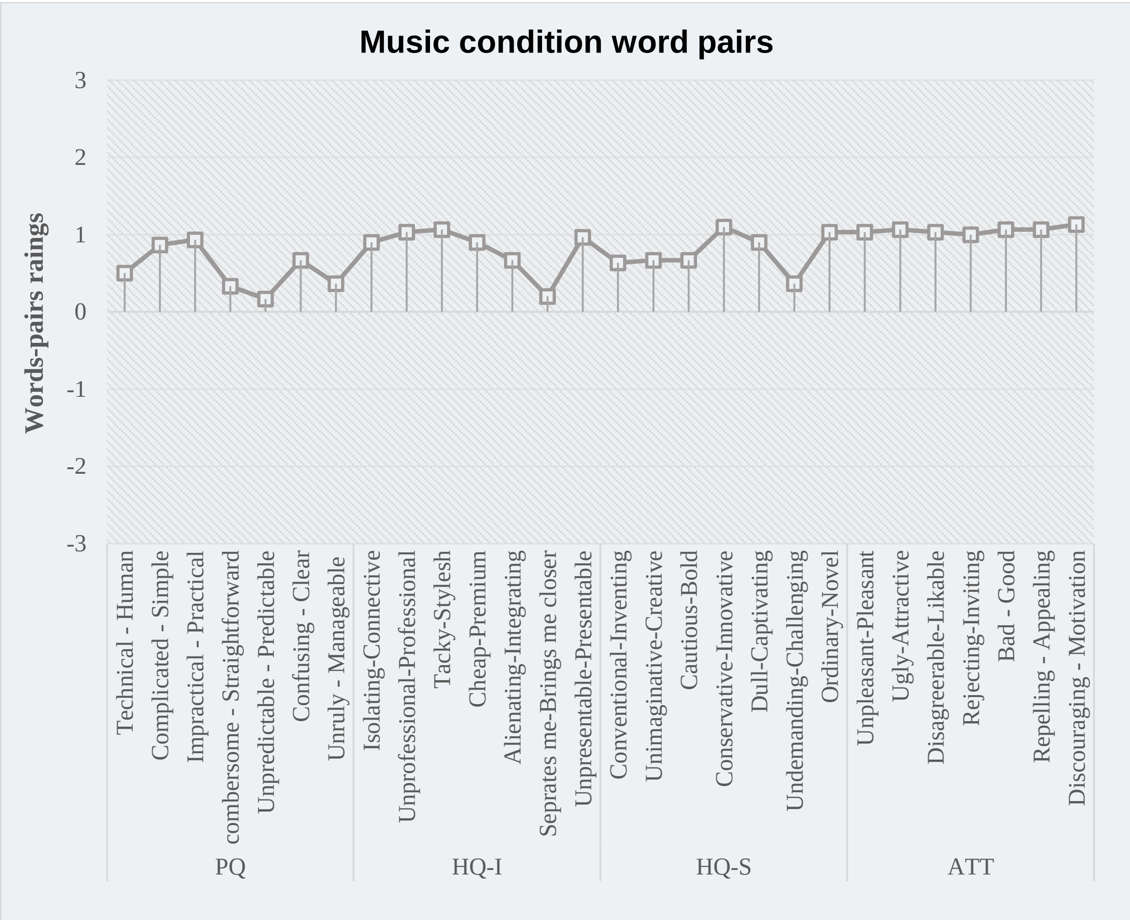


Figure 5-12: AttrakDiff word-pairs (music condition).

With regards to the classical music condition, the condition perceived PQ, to some extent users connected well with this condition as they considered it to be less technical, their rating was above average, and they thought it was simple, practical and straightforward. Concerning HQ-I, users rated the condition as connective, professional, stylish, integrating and presentable. In relation to HQ-S, users were enthused by the condition; they considered it to be creative, bold, innovative and pleasant. In terms of condition attractiveness, it was rated above average, considered to be pleasant, likable, tempting, useful and inspiring.

5.4.5 **Speech Condition User Experience**

The speech condition was ranked by the users as "task-oriented". Pragmatic quality is precisely the classification. It is pragmatic and the condition optimally assisted its users. The pragmatic quality was moderate; therefore, there is possible room for enhancements within this category (see Figure 5-13). In terms of hedonic quality, the character classification applies positively. The users were encouraged by the condition; however, the hedonic value is only above the average value, subsequently there is opportunity for enhancements. The confidence rectangle is small within the pragmatic dimension and stretched within the hedonic side; users are at one in their evaluation of the pragmatic quality while they were in less agreement on their rating of the hedonic quality.

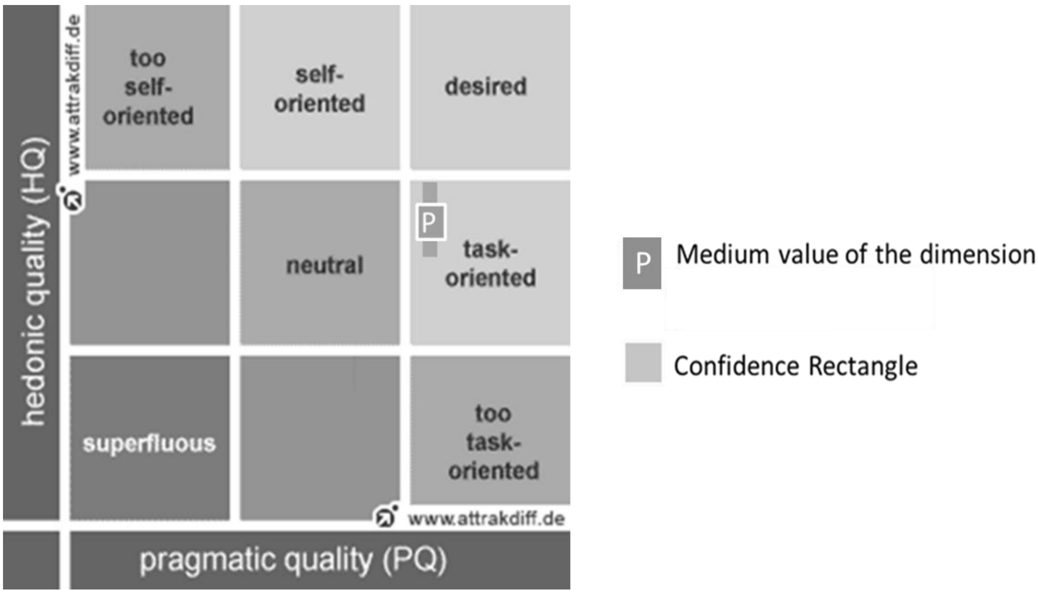


Figure 5-13: PQ and HQ medium values (speech condition).

5.4.5.1 ***Mean values of users' experience***

The mean values of the AttrakDiff evaluationdimensions for the condition are shown in Figure 5-14 in which, the stimulation and identity aspects of hedonic qualities are well distinguished. Furthermore, the rating of attractiveness is presented.

Hedonic quality – identity: This condition is placed in the mean region. It provides the user with identification and thus meets ordinary standards. To bind the user more strongly to the speech condition, improvement is necessary.

Hedonic quality – stimulation: This condition is positioned above the average region. It meets normal standards; to inspire and stimulate users, additional improvements are needed.

Attractiveness: The condition's attractiveness score is placed in the middle region. The overall impression of this condition is moderately attractive.

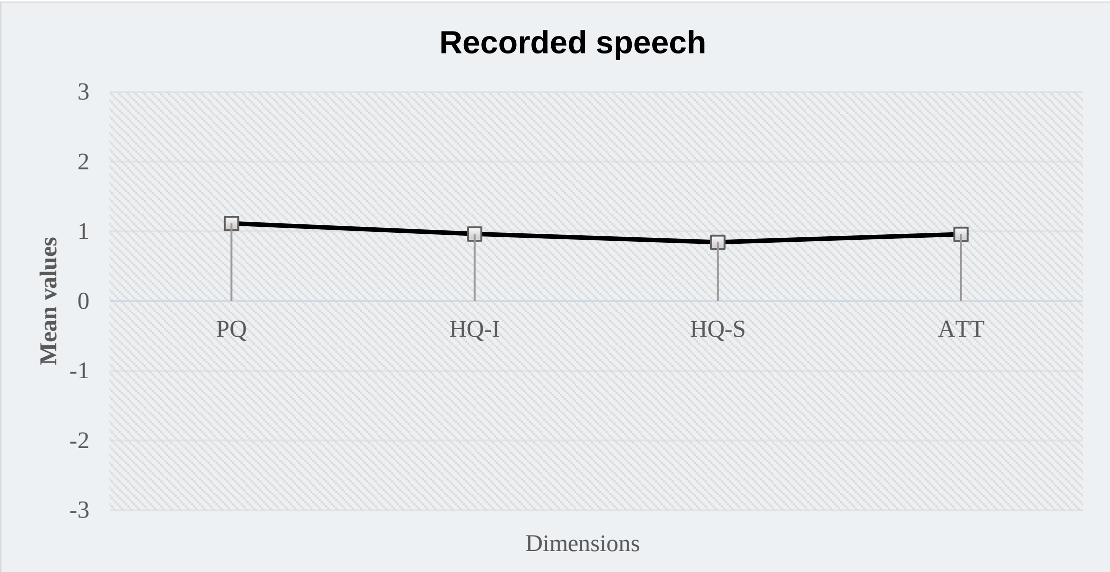


Figure 5-14: AttrakDiff dimensions' mean values (speech condition).

5.4.5.2 *Description of Word-Pairs*

The mean values of the word pairs are presented in (see Figure 5-15), The results show which characteristics are particularly critical or particularly well resolved With respect to the speech condition's perceived PQ, to some degree users connected well with this condition, their rating was above average, they thought it was simple, practical, clear, straightforward and manageable. Concerning HQ-I, users thought the condition was professional, to some degree stylish, premium and presentable. Concerning HQ-S, users were stimulated by the condition; they considered it to be creative, bold, innovative and pleasant. In terms of condition attractiveness, it was rated above average; it was considered to be pleasant, likable and useful.

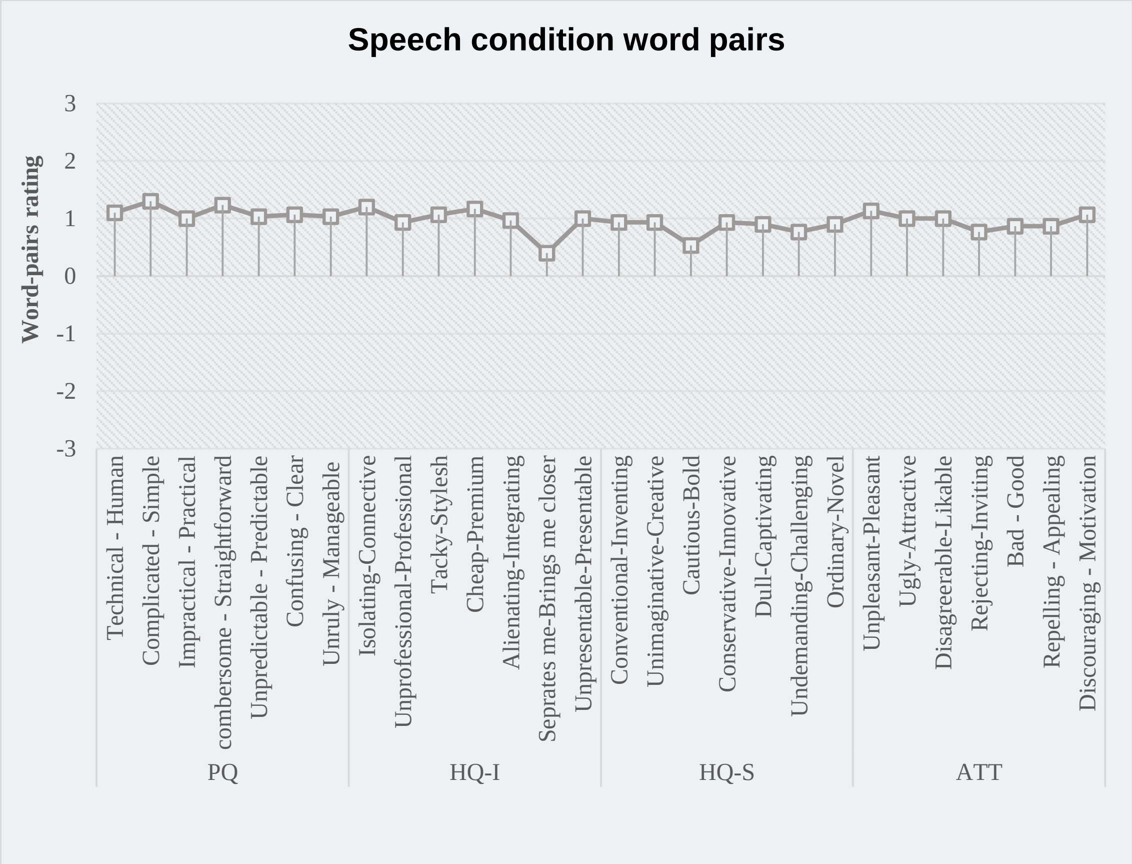


Figure 5-15: AttrakDiff word-pairs (speech condition).

5.4.6 NMMC (Controlled) Condition User Experience

The control condition in terms of pragmatic quality was valued “neutral”. In terms of hedonic quality, it was deemed “superfluous”.

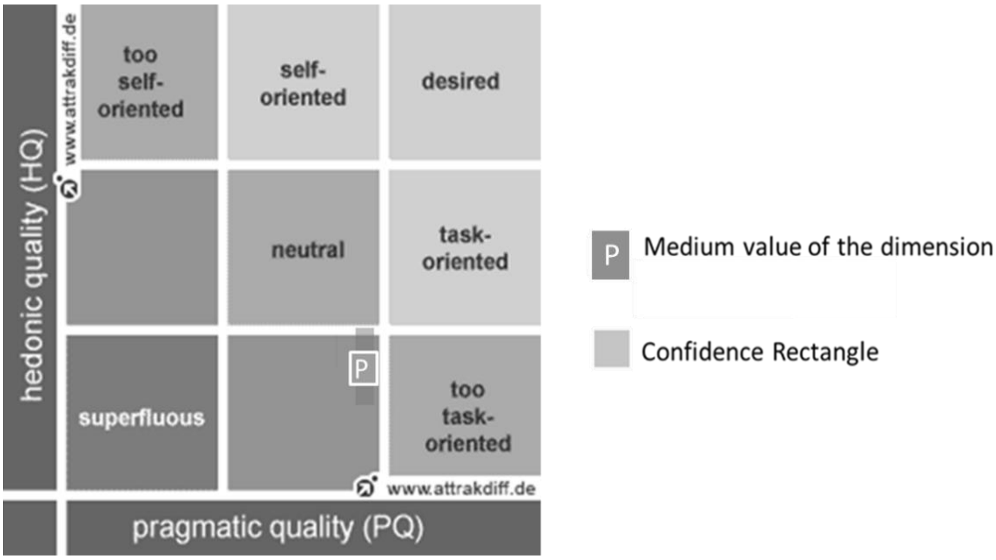


Figure 5-16: PQ and HQ medium values (NMMC condition).

The users were assisted by it, however the value of pragmatic quality only reaches average value; consequently there is a chance for large improvements improvement within this category terms of usability (see Figure 5-16). Concerning the hedonic quality, the character classification clearly does not apply because the confidence interval spills out over the character zone. The rating within this dimension was poor and users could identify with it nor were they inspired by it. Clearly, there is an urgent for this condition to this condition to undergo a makeover. The confidence rectangle is small, that indicates users were in less agreement on their assessment of hedonic quality than what they did for pragmatic quality.

5.4.6.1 *NMMC Mean values of users experience*

The means values of the AttrakDiff dimensions for the condition are shown in Figure 5-17, in which the stimulation and identity aspects of hedonic qualities are distinguished between. Furthermore, the rating of attractiveness is presented.

Pragmatic quality: The condition score within this dimension was its best performance area, it was deemed above the average, however still there is a large room for improvements

Hedonic quality – identity: This condition is placed below the average region and t struggles to meets the ordinary standards. To bind the user to the condition, improvement must be made.

Hedonic quality – stimulation: This condition is placed way below average value. It does not have a stimulating effect on users. Inadequate motivation results in a lack of drive to use the condition. This condition is in urgent need of revision.

Attractiveness: This condition’s attractiveness value is placed in the below-average region; the whole impression of the condition is unattractive.

5.4.6.2 *Description of Word-Pairs*

The mean values of the word pairs are presented here. The results show which characteristics are particularly in need of urgent attention or particularly well resolved (see Figure 5-18). Concerning NMMC condition’s PQ, users rated them above average with one exception were it was viewed to be more technical than human; they thought it was simple, practical, clear and manageable. With regards HQ-I, users thought the condition was isolating, to some point tacky, cheap and to some extent it was considered to be presentable. Concerning HQ-S, users were not stimulated by the condition; they considered it to be extremely conventional, cautious, unadventurous and ordinary. With respect to the conditions' attractiveness, it was rated below average, considered ordinary, disagreeable, ugly and discouraging.

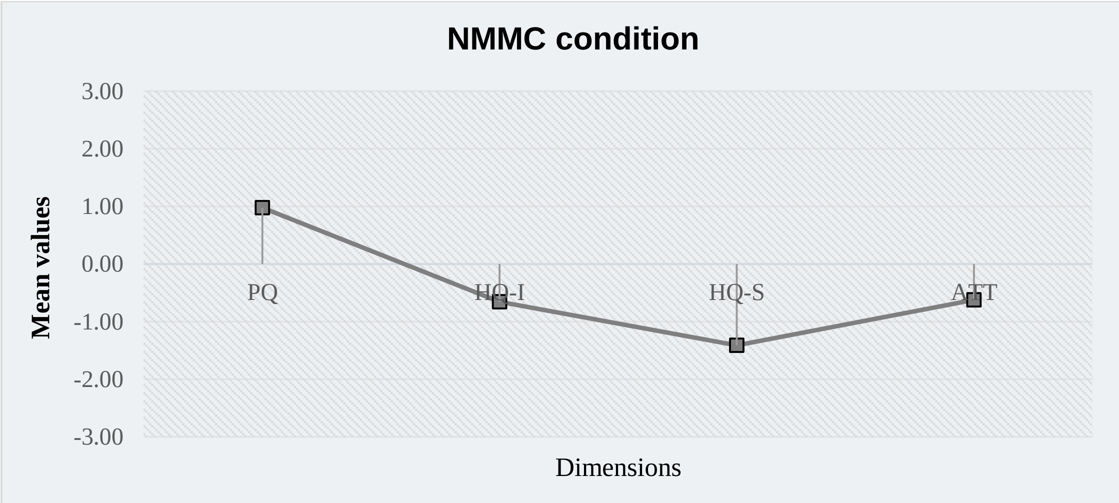


Figure 5-17: AttrakDiff dimensions' mean values (NMMC condition).

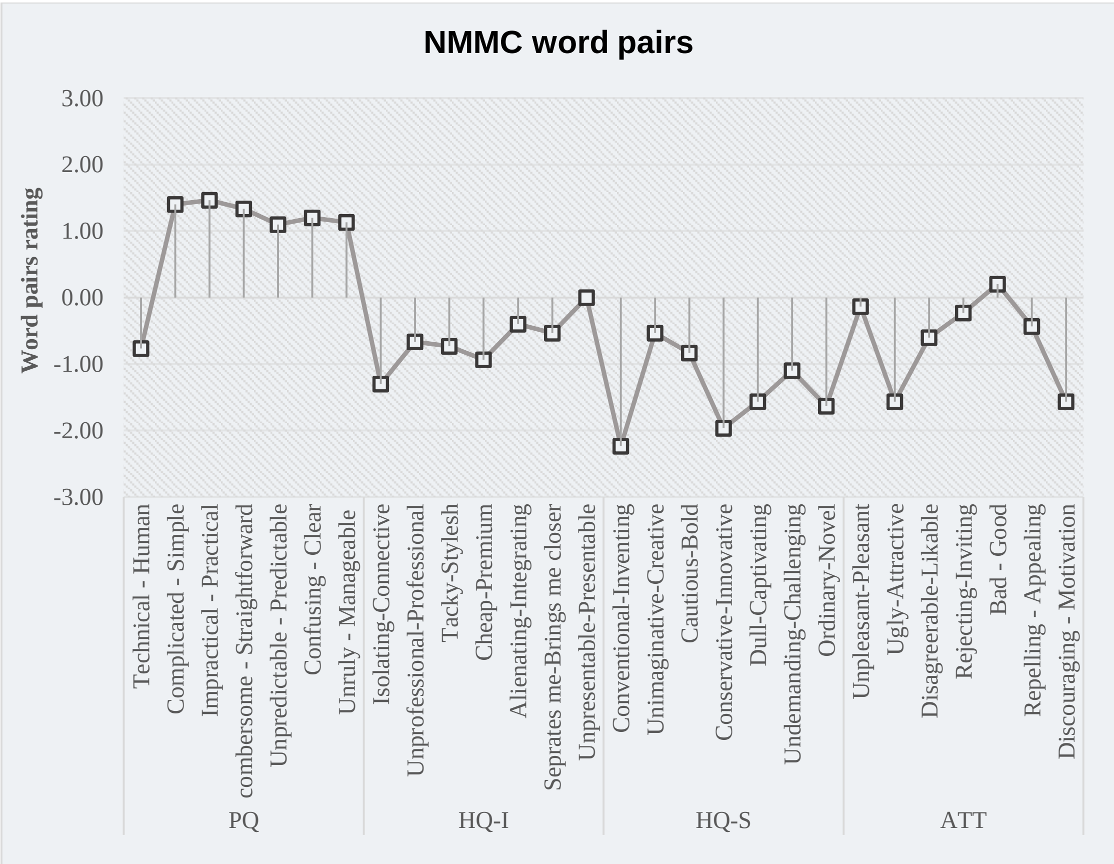


Figure 5-18: AttrakDiff word-pairs (NMMC condition).

5.4.7 **Statistical Data Analysis**

This analysis aims to find out whether the perceived differences among the tested conditions in users' experience have any statistical significance. To select an appropriate statistical test, the data was tested to find if it was normally distributed.

5.4.7.1 ***Normality Test***

The user experience data was tested for normality using the "Shapiro-Wilk" test (see Table 5-1). The results obtained show that not all data was normally distributed, taking into account the following:

- The study uses a repeated within-subjects measurements
- The data to be compared is not normally distributed

- Related samples Friedman's two way analysis of variance (ANOVA) by ranks test, which is a non-parametric test was used to measure the differences between the tested conditions in terms of user experience.

Tests of normality						
Condition	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.102	30	.200[*]	.973	30	.618
Avatar	.158	30	.055	.897	30	.007
Earcons	.115	30	.200[*]	.955	30	.228
Music	.188	30	.008	.888	30	.004
Speech	.159	30	.051	.933	30	.058
NMMC	.115	30	.200[*]	.952	30	.189
*. This is a lower bound of the true significance.						
a. Lilliefors Significance Correction						

Table 5-1: Users' experience normality test.

5.4.7.2 ***Friedman's ANOVA Test***

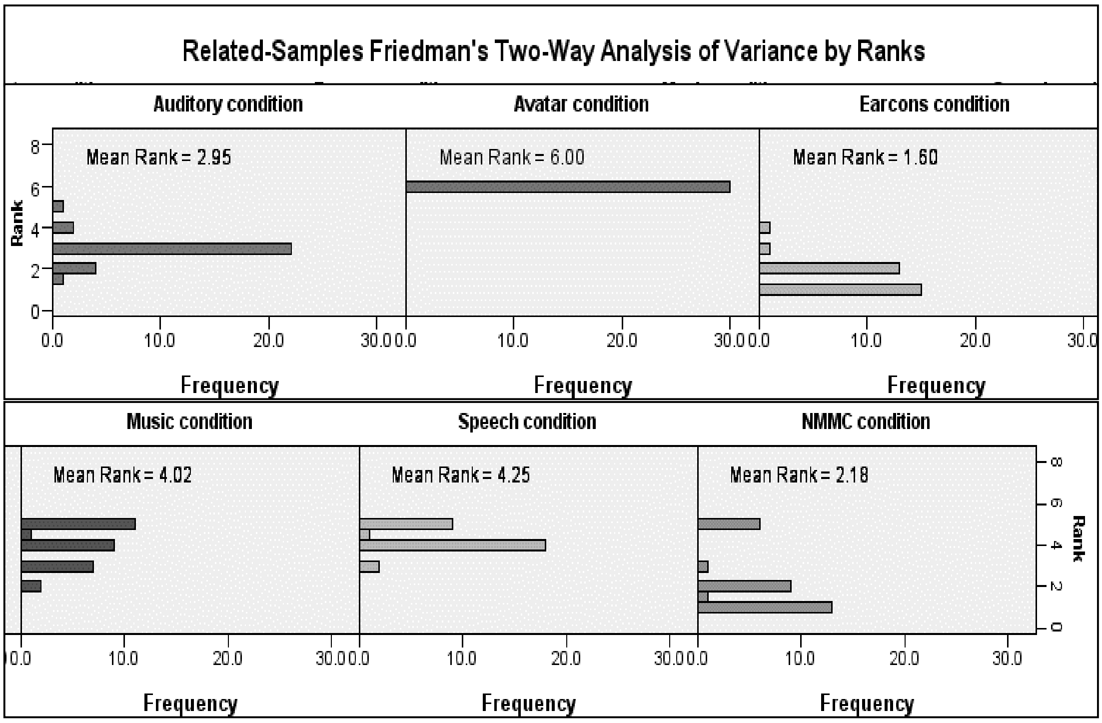
There is a null hypothesis (H0) and an alternative hypothesis (HA) to test and they are stated below:

$$H0: \mu_{Auditory}=\mu_{Avatar}=\mu_{Earcons}=\mu_{Music}=\mu_{NMMC}=\mu_{Speech}$$

$$HA: \mu_{Auditory}\neq\mu_{Avatar}\neq\mu_{Earcons}\neq\mu_{Music}\neq\mu_{NMMC}\neq\mu_{Speech}$$

The Friedman's test result ($\chi2\ (5) = 111.846$, $p < .0005$.) shows that the difference between conditions in their provided user experience is statistically significant. Therefore, the null hypothesis is rejected and the alternative hypothesis is accepted (for detailed test results see Figure 20).

The Freidman's ANOVA test did not show where the difference lies so to find that, a *post-hoc* (pairwise comparison) test is needed. The test was executed and the results are presented in Figure 5-20, the significant differences are denoted in gray.



Hypothesis Test Summary

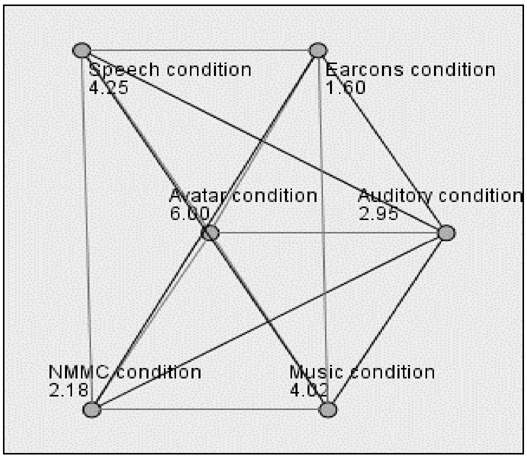
	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory condition, Avatar condition, Earcons condition, Music condition, Speech condition and NMMC condition are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	111.846
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 5-19: User experience Freidman's ANOVA test.

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Earcons condition-NMMC condition	-.583	.483	-1.208	.227	1.000
Earcons condition-Auditory condition	1.350	.483	2.795	.005	.078
Earcons condition-Music condition	-2.417	.483	-5.003	.000	.000
Earcons condition-Speech condition	-2.650	.483	-5.486	.000	.000
Earcons condition-Avatar condition	4.400	.483	9.109	.000	.000
NMMC condition-Auditory condition	.767	.483	1.587	.112	1.000
NMMC condition-Music condition	1.833	.483	3.795	.000	.002
NMMC condition-Speech condition	2.067	.483	4.278	.000	.000
NMMC condition-Avatar condition	3.817	.483	7.901	.000	.000
Auditory condition-Music condition	-1.067	.483	-2.208	.027	.408
Auditory condition-Speech condition	-1.300	.483	-2.691	.007	.107
Auditory condition-Avatar condition	-3.050	.483	-6.314	.000	.000
Music condition-Speech condition	-.233	.483	-.483	.629	1.000
Music condition-Avatar condition	1.983	.483	4.106	.000	.001
Speech condition-Avatar condition	1.750	.483	3.623	.000	.004

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 5-20: User experience pairwise comparison.

In order to understand the differences perceived in pairwise comparison between the tested conditions in the quality of their provided user experience, the conditions are ranked accordingly in Figure 5-21. The most significant difference was observed between the avatar condition, followed by the speech and music conditions. The least differences were noticed between the earcons condition, the auditory condition, and the NMMC condition. Furthermore, the figure shows that, in terms of user experience the highest rank was for the avatar condition followed by the speech condition, then the music condition, then the auditory condition and then the NMMC. The lowest rank in terms of user experience was for the earcons condition, which was the only condition ranked below the control condition.

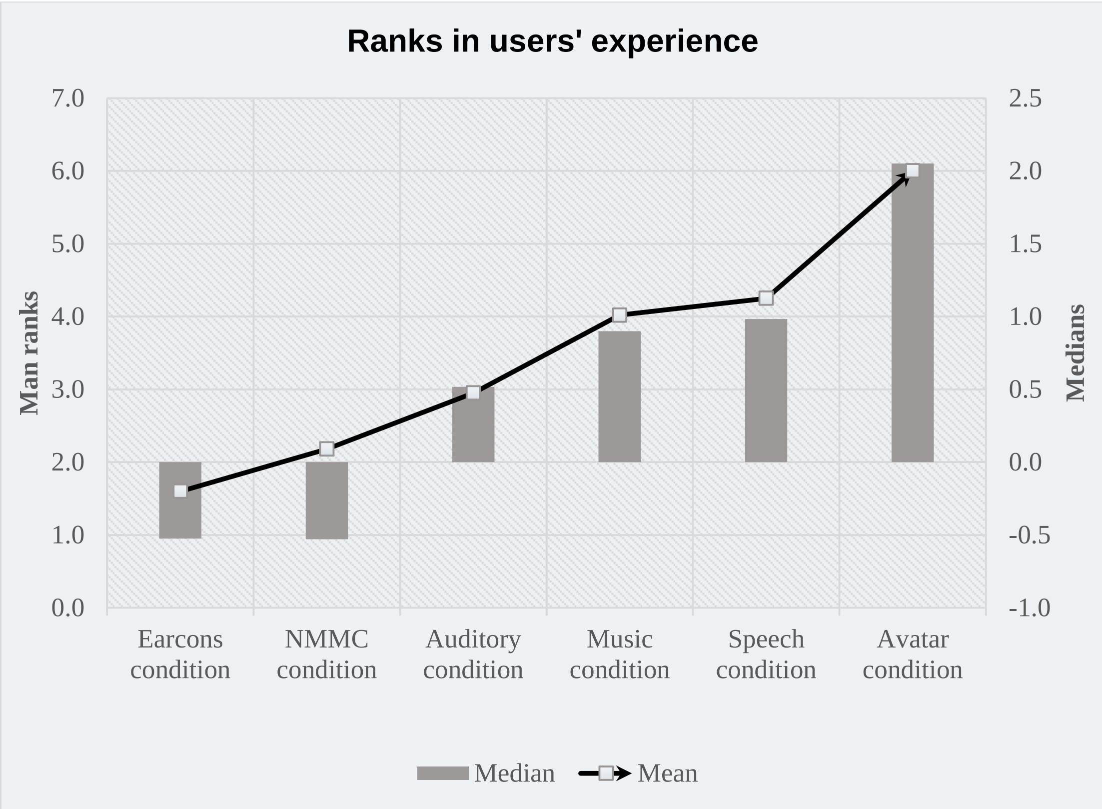


Figure 5-21: User experience (pairwise comparison) ranks.

5.5 DISCUSSION

To assess the impact of each individual communication metaphor used in the experiment on the overall user experience and how each one was perceived by the users, the results obtained in each experimented condition are discussed in comparison with the control condition (NMMC) and other experimented conditions.

5.5.1 Auditory Icon Condition vs. NMMC Condition

Figure 5-22 shows both auditory and control condition presented against each other. Overall, users had a positive view of the auditory icon condition; the noticed improvements over the control condition can be attributed on to the use of auditory icons communication metaphor. With reference to the results obtained in Figure 5-3, they considered it to be pragmatically above average. To some extent, users related to this condition, they rated it as to some extent simple, practical, clear and manageable. In terms of hedonic qualities (HQ-I) and (HQ-S) they were stimulated and identified with it. They thought the condition was professional, to some extent stylish, premium and presentable.

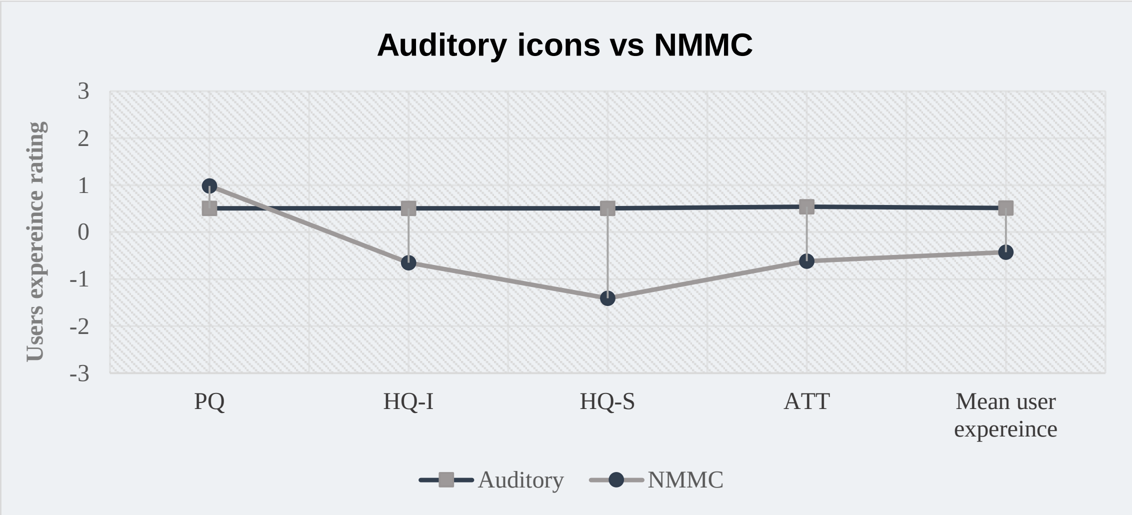


Figure 5-22: Attrakdiff (auditory icons vs NMMC).

Users were slightly enthused by the condition; they considered it innovative, bold, and slightly pleasant. In terms of condition attractiveness, it was rated slightly above average, considered agreeable, useful and they were not discouraged by using it. With regard to attractiveness (ATT), users found it somewhat attractive. Overall the auditory icons surpassed the NMMC in terms of hedonic (HQ-I, HQ-S) quality and attractiveness. The only exception was in the pragmatic quality class where it was rated as less pragmatic. It seems the use of auditory icons improved the condition's hedonic value and made it more attractive and appealing to the users.

5.5.1.1 *Avatar condition vs NMMC condition*

The results from the avatar condition is depicted, together with those obtained from the control condition, in Figure 5-23 the results clearly show that the experimental condition surpassed the control condition within all of the evaluation dimensions and the observed improvements can only be attributed to employment of the avatar interaction metaphor. With reference to the results in Figures 5-4, 5-5 and 5-6 the avatar condition was positively rated by all of the users and they considered it highly pragmatic. Users related very well to the avatar condition, as they believed it was more human than technical, their ranking was high and they assumed it was modest, convenient, exciting and manageable.

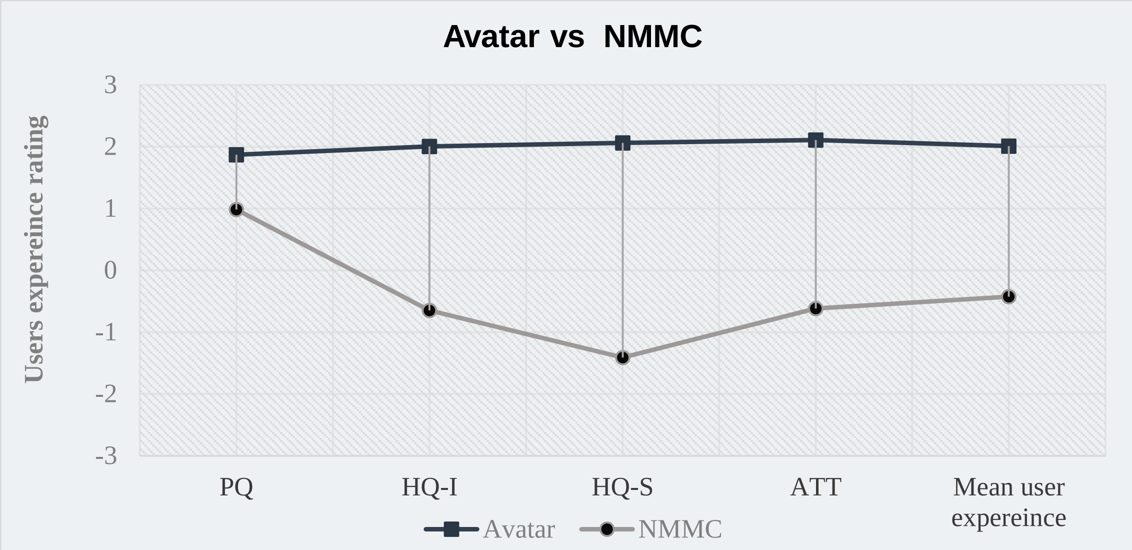


Figure 5-23: AttrakDiff (avatar vs NMMC condition).

Concerning the condition's hedonic quality, users viewed the condition with great interest and they considered it proficient, very elegant and exceptionally neat. Furthermore, users were highly enthused by it; they considered it innovative, courageous, appealing, novel and pleasant. In relation to the condition's perceived attractiveness, user rated it as vastly attractive, fair, friendly, valuable and encouraging.

5.5.2 Earcons Condition vs. NMMC Condition

Figure 5-24 shows a comparison between the earcons condition and the control condition. The earcons condition was rated less positive than the control condition. With reference to the results presented in Figures 5-7, 5-8 and 5-9 in terms of pragmatic quality, users found it to be more technical and complicated. With respects to the hedonic quality identification, users rated it below average, users were pessimistic in their rating; they considered it more complicated and disruptive. Furthermore, users did not identify themselves well with the earcons condition as they felt it was tacky, cheap, alienating and unsuitable, it is worth mentioning here that in this dimension it was rated relatively similar to the control condition. In terms of hedonic qualities-stimulation, they were somehow stimulated by it; they thought it was bold and in this category, it performed better than the NMMC condition. Concerning the condition attractiveness (ATT), users thought it was marginally less attractive than the control condition; they felt it was rejecting, ugly and ordinary. Overall, there were no noticeable improvements resulting from the inclusion of the earcons metaphor except for its ability to stimulate users; on the other hand, it hindered the users' experience in other areas and therefore it should be used cautiously.

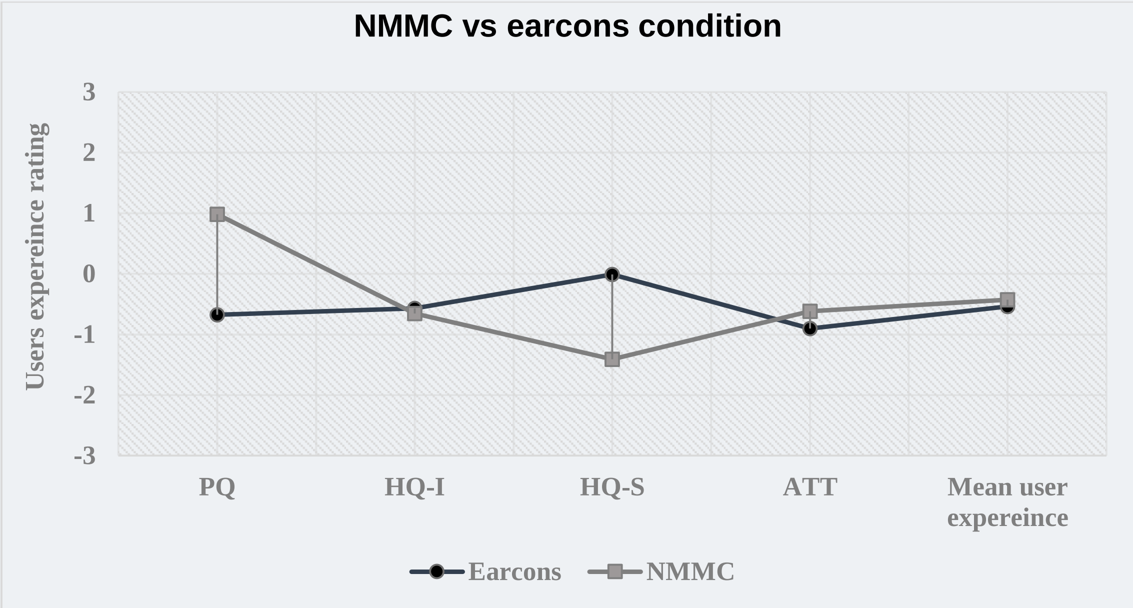


Figure 5-24: AttrakDiff (earcons vs NMMC).

5.5.3 Music Condition vs NMMC Condition

For the music condition, the comparison with the control condition is portrayed in Figure 5-25. In general, users had a more positive view of the music condition and the only downside to it was in terms of pragmatic quality as it was equal to the NMMC. With reference to Figures 5-10, 5-11 and 5-12; the classic music condition professed moderate pragmatic quality (PQ), users associated well with it; they considered it to be simple, practical, straightforward, less technical and more human. Concerning hedonic quality identification (HQ-I), reflecting on their experience users regarded the condition as connective, professional, tasteful, constructive and well turned-out. In relation to hedonic quality stimulation (HQ-S), users were inspired by the condition; they considered it artistic, courageous, pioneering and lovely. With respect to the condition attractiveness, it was regarded to be enjoyable, easy-going, inviting, valuable and exciting.

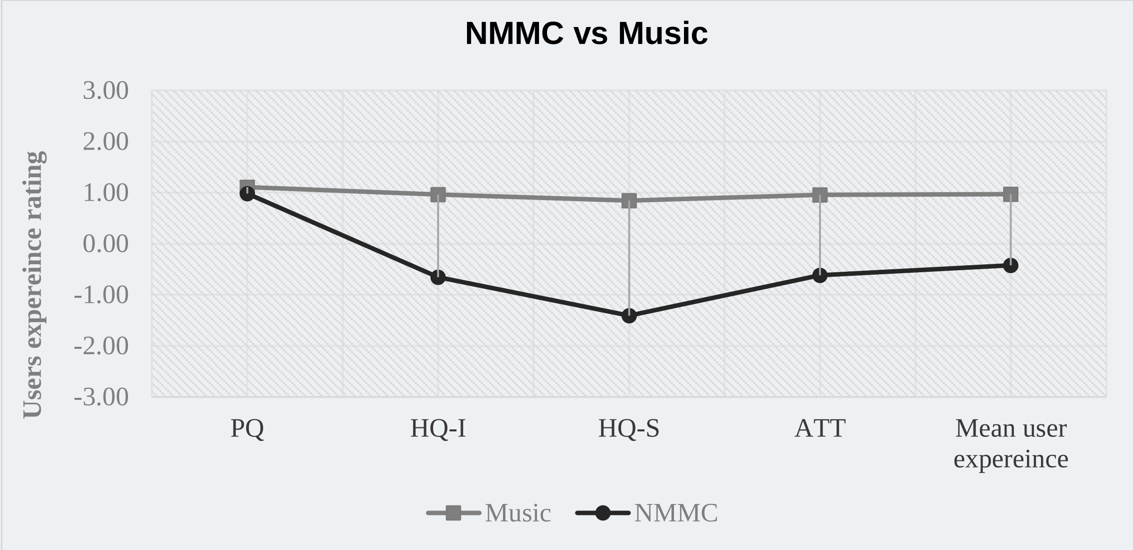


Figure 5-25: AttrakDiff (music vs NMMC).

Overall, the results obtained from this experimental work clearly indicate that the inclusion of classical music led to a more optimistic user experience in comparison with the control condition. The greatest improvement was the added value to the system's hedonic quality and attractiveness.

5.5.4 **Speech condition vs. NMMC Condition**

Figure 5-26 presents a comparison between the speech condition and the control condition; overall, users were more positive reflecting on their experience with speech condition; they valued both to be equally pragmatic.

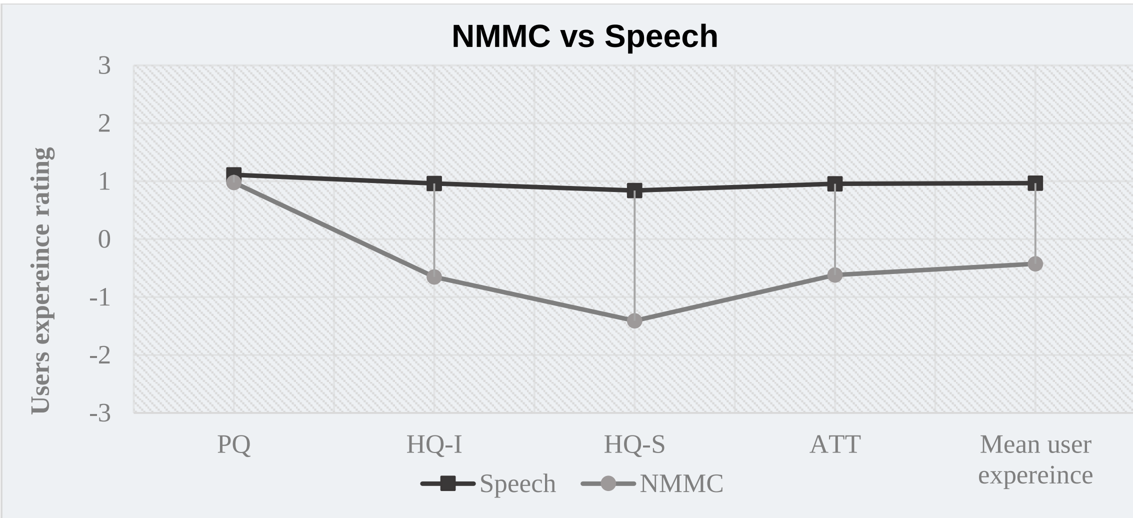


Figure 5-26: AttrakDiff (speech vs NMMC).

However, there was a more optimistic view for the speech condition within the other user experience dimensions.

Looking at Figures 5-13, 5-14 and 5-15; in relation to the speech condition's apparent pragmatic quality (PQ), to some extent users related well with this condition, their rating was beyond average, they considered it to be modest, practical, flawless, straightforward and handy. Concerning hedonic quality identification (HQ-I), users identified themselves well with it; they considered the condition to be professional, to some degree classy and presentable. With reference to hedonic quality stimulation (HQ-S), users were encouraged by the condition; they regarded it to be resourceful, useful and pleasant. In terms of condition attractiveness (ATT), it was valued to be moderately attractive, it was deemed pleasant, friendly and advantageous. Overall, the inclusion of the speech interaction metaphor moderately improved users' experience in comparison with the control condition. From the observed user experience improvements seen within the experimental condition in comparison with the control condition, it is safe to conclude that, employing speech modality could provide added-value to e-learning interactive systems by improving the learner's user experience.

5.5.5 All Conditions

Figure 5-27 presents the results of all the conditions' evaluations side by side. Generally, the most prominent improvement was in the avatar condition followed by the music and speech conditions, which provided relatively similar user experiences, and then the auditory icons condition, which was more positive than the control condition; the least improvement of all was in the earcons condition. The downfall of the NMMC was its lack of hedonic quality attributes and inability to attract the users. With regards to the earcons condition the users found it stimulating, however over-use of the interaction metaphor hindered other dimensions of the users' experience as they were neither attracted to it nor able to identify with it and therefore it should be used carefully. The results of user experience evaluation revealed the importance of the system's hedonic quality values and its attractiveness in gaining users' attention and acceptance. Furthermore, the evaluations of user experience tapped into important areas of the relation between the user and the system, which were not covered by the conventional usability evaluation method, which merely investigated the functionality and pragmatic side of the system.

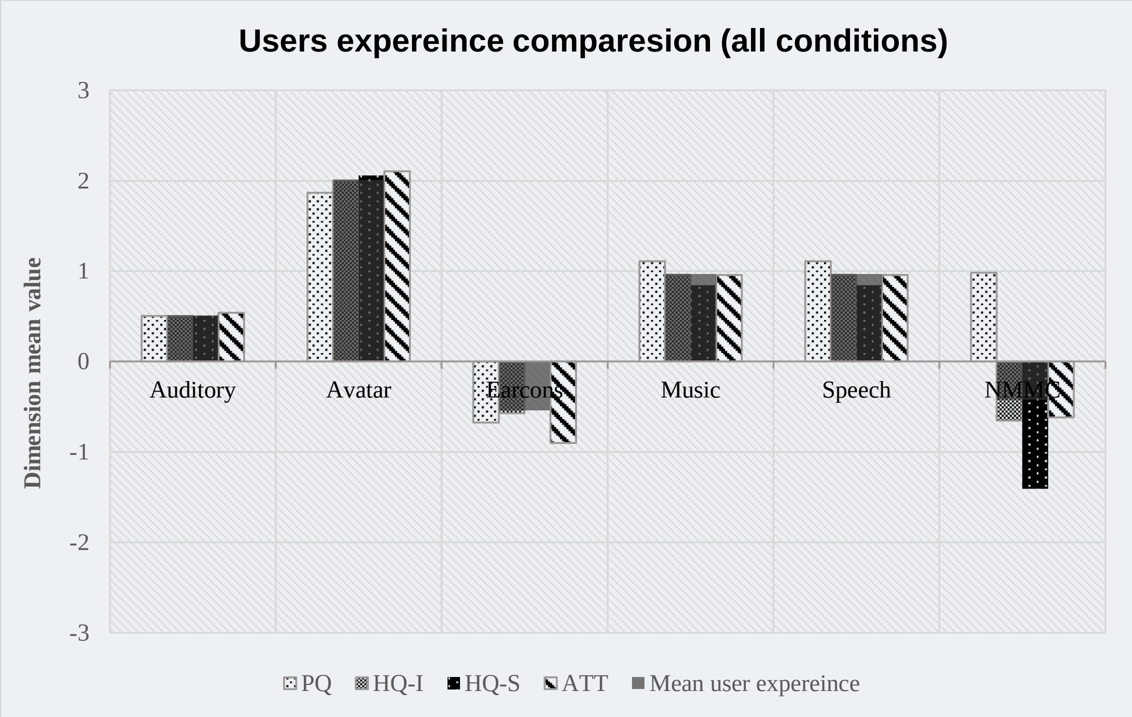


Figure 5-27: All conditions user experience comparison.

User experience evaluation revealed what users really valued most in dealing and interacting with the system. The empirical results obtained in this experiment clearly show the importance of using this evaluation method side by side with conventional usability evaluation criteria in assessing e-learning interactive systems. Based on the aforementioned results of this experimental work and performed data analysis:

H1: This hypothesis was rejected, as the experimental conditions did not receive equal preferences and rating in their provided pragmatic user experience measurements.

H2: This hypothesis was rejected as the experimental work proved that conditions did not receive equal preferences in terms of users' identification with conditions' provided hedonic user experience.

H3: The hypothesis was overruled based on the obtained results, as the experimental conditions did not equally stimulate the users with their provided hedonic users' experience.

H4: The experimental conditions were not equally appealing and attractive in their provided user experience; therefore, it was proven wrong.

5.6 SUMMARY

A within subjects measurements study was conducted for the assessment of user experience of an e-learning platform in which five different experimental conditions and one control condition were evaluated. Each experimental condition was dominated by a single interaction metaphor (auditory icons, facially expressive speaking avatar, earcons, music, recorded speech) and the aim was to find out how each metaphor could influence the overall user experience. The evaluation method was subjective in nature; users were requested to use AttrakDiff instrument to reflect on and report their experiences with each of the six conditions. The best user experience was observed when using the avatar communication metaphor, followed by the recorded speech and the music metaphors and then the auditory icons. The worst user experience was observed with using earcons communication metaphor. An overview of the users' experience evaluation results is presented in the user experience traffic light in Figure 5-28.

User experience traffic light					
Condition	**PQ**	**HQ-I**	**HQ-S**	**ATT**	**UX**
Auditory	!	✗	!	!	!
Avatar	✓	✓	✓	✓	✓
Earcons	!	✗	!	✗	✗
Music	✓	✓	✓	✓	✓
Speech	✓	✓	✓	✓	✓
NMMC	!	!	!	!	!

Figure 5-28: The user experience evaluation traffic light.

CHAPTER 6

6 AFFECTIVE STATE EVALUATION (EXPERIMENTAL PHASE II)

Users' affective state evaluation differs greatly from usability (Chapter 4) as well as user experience(Chapter 5) evaluations because it aims at revealing the influence each experimental conditions has on users' emotion and how they positively or negatively are moved by the employed communication metaphors. User affective state has two components, physiological and psychological, accordingly the study adapted subjective (SAM) and objective (Biofeedback) evaluation techniques to measure both components based on the PAD [101] affective state model.

6.1 AIMS and OBJECTIVES

The main aim of this evaluation is to examine the impact of each communication metaphor on user's affective state within three dimensions (Valence, Arousal and Dominance). To accomplish the aims of this evaluation, these objectives have to be met:

- Formulate experimental hypotheses.
- Evaluate users affective state through utilizing self-assessment manikin technique and biofeedback (skin conductance and temperature) measurements with the following conditions:
    - Auditory condition.
    - Music condition.
    - Earcons condition.
    - Avatar condition.
    - Speech condition.
    - Control condition.

6.2 HYPOTHESES

It is anticipated that making use of communication metaphors will equally influence users' affective state and the following hypotheses have been formulated accordingly:

H1: The experimental conditions will induce an equal positive valence.

H2: The experimental conditions will be equally arousing.

H3: The experimental conditions will be equally dominating.

H4: The experimental conditions will induce the same level of skin conductance.

H5: The experimental conditions will trigger the same level of skin temperature.

H6: The experimental conditions will induce an equal affective state.

6.3 METHOD OF INVESTIGATION

Users' emotions towards the experimented conditions in terms of pleasure, arousal and dominance were evaluated using two techniques, objective assessment using biofeedback measurements (skin conductance and temperature) in addition to Self-Assessment Manikin (SAM) techniques. SAM is a non-verbal assessment tool, which is cross-cultural and easy to administer and understand. Upon experiencing each condition, users were requested to use SAM technique to report their affective state and the results are presented here accordingly.

6.4 SELF-ASSESSMENT MANIKIN (SAM) TECHNIQUE

6.4.1 SAM (Auditory Icons) Condition Results

Users' assesment of the auditory icons condition within the valence dimension was "slightly negative" emotion towards the condition.

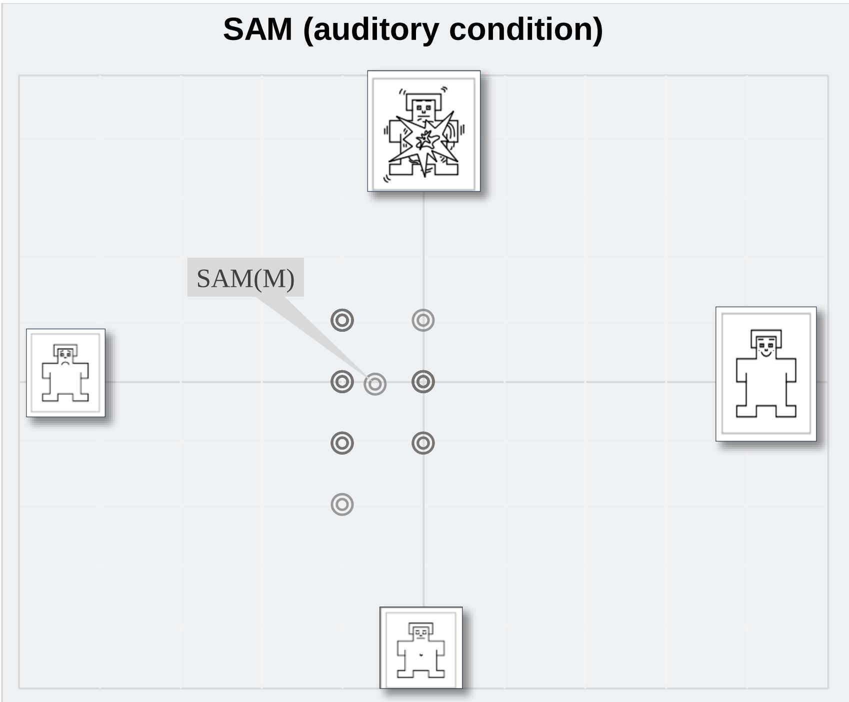


Figure 6-1: SAM result (auditory icons condition) results.

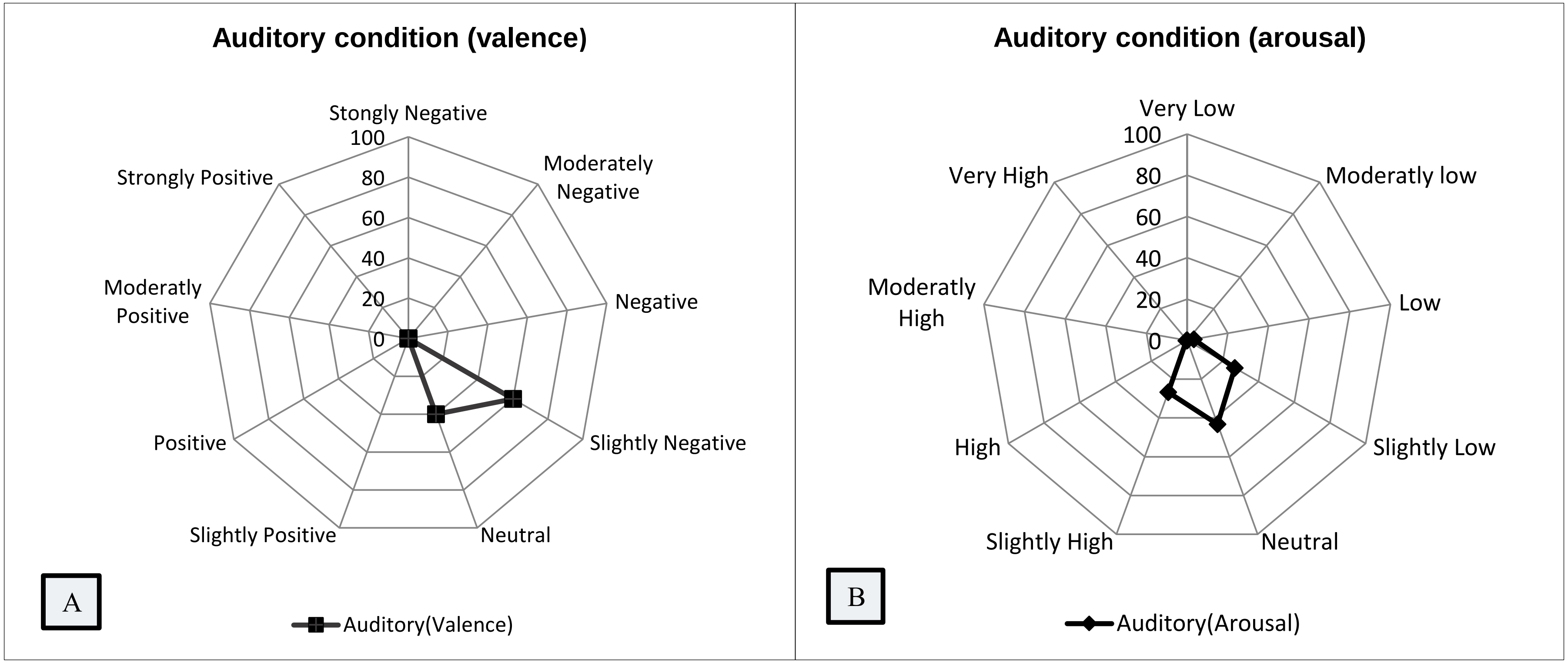


Figure 6-2: Auditory icons condition's radar plots.

in the arousal dimension users' reaction was considered to be "neutral" and the overall impression of the auditory condition was "neutral" (see Figure 6-1). For a better understanding of user ratings in both dimensions the results are plotted in a radar chart (see Figure 6-2 A,B); user ratings for the valence dimension is moving from "neutral" towards "slightly negative", while in the arousal dimension users' rating was placed between "slightly low" and "slightly high" with a tendency towards neutral.

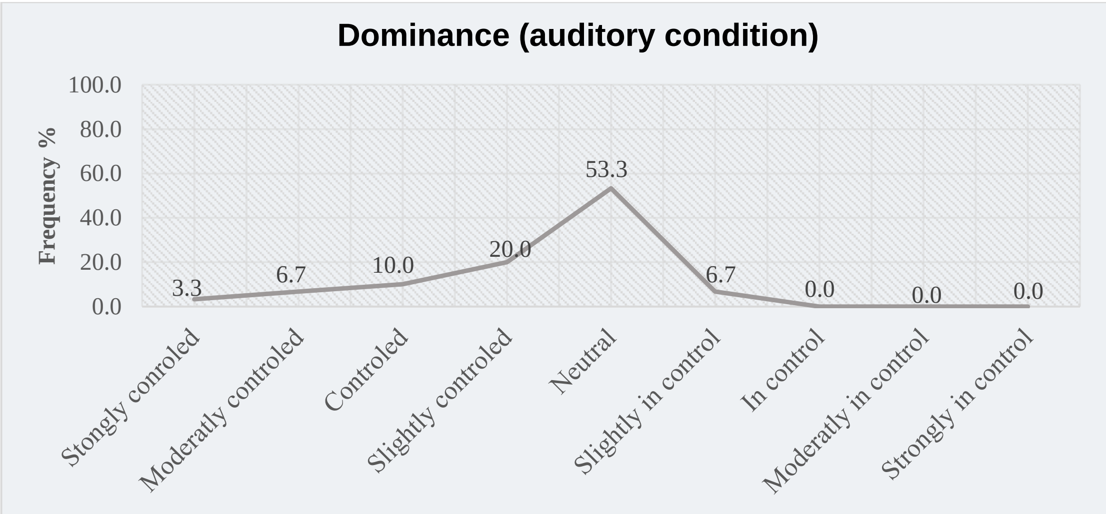


Figure 6-3: Dominance (auditory condition) results.

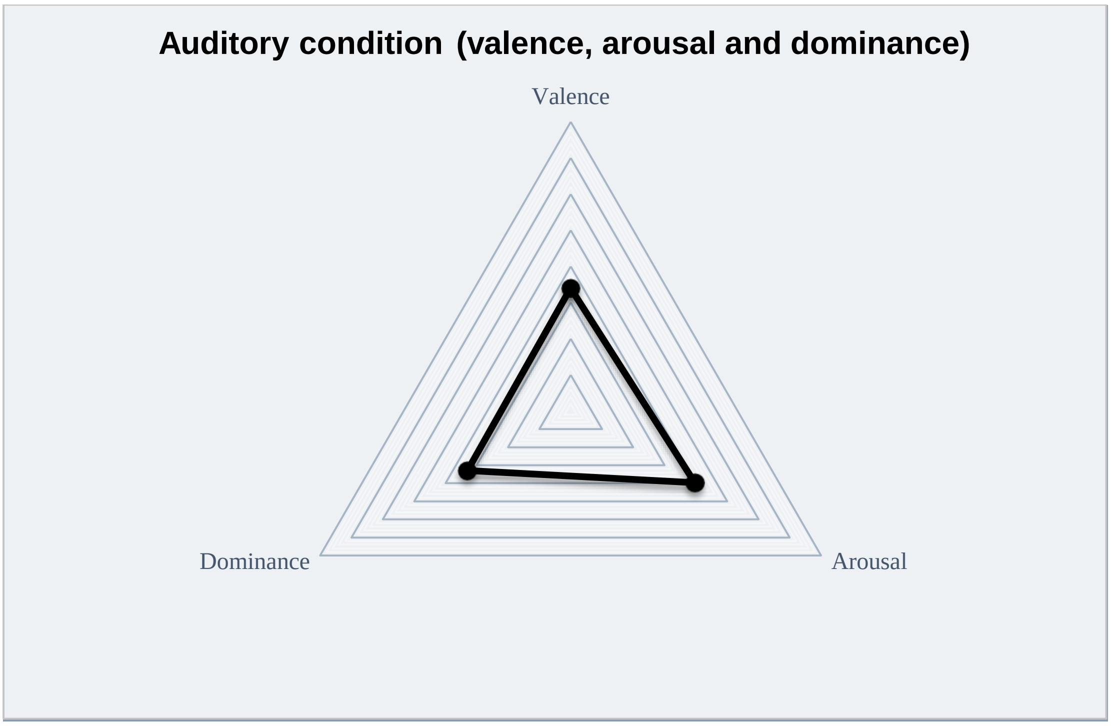


Figure 6-4: Auditory condition (valence, arousal and dominance).

The areas covered with line graphs is small which indicates that there is considerable agreement among users in their rating. Concerning the dominance dimension (see Figure 6-3) users ratings in general were “neutral” with a tendency towards feeling “controlled”. A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users SAM dimensions. There was a weak negative correlation between the two dimensions ($r_s(28) = -.213$, $p = .257$). The overall impression is that users felt the condition was slightly negative and to some extent, they were “in control”. The layout of the line graph in Figure 6-4 is moving towards the arousal angle; this indicates that users were moved by the auditory condition.

6.4.2 SAM (Avatar Condition)

User ratings for the valence and arousal dimensions are shown in Figure 6-5. Users’ assesment of the condition was “positive” and within the arousal dimension users’ reaction was “neutral” and the overall impression of the auditory condition was “slightly high”. The results of user ratings in both dimensions are plotted in a radar chart (see Figure 6-6 (A,B)); user ratings for the valence dimension are moving towards “moderately positive”, while in the arousal dimension user ratings were placed between “slightly low“ and ”slightly high” with a tendency towards neutral.

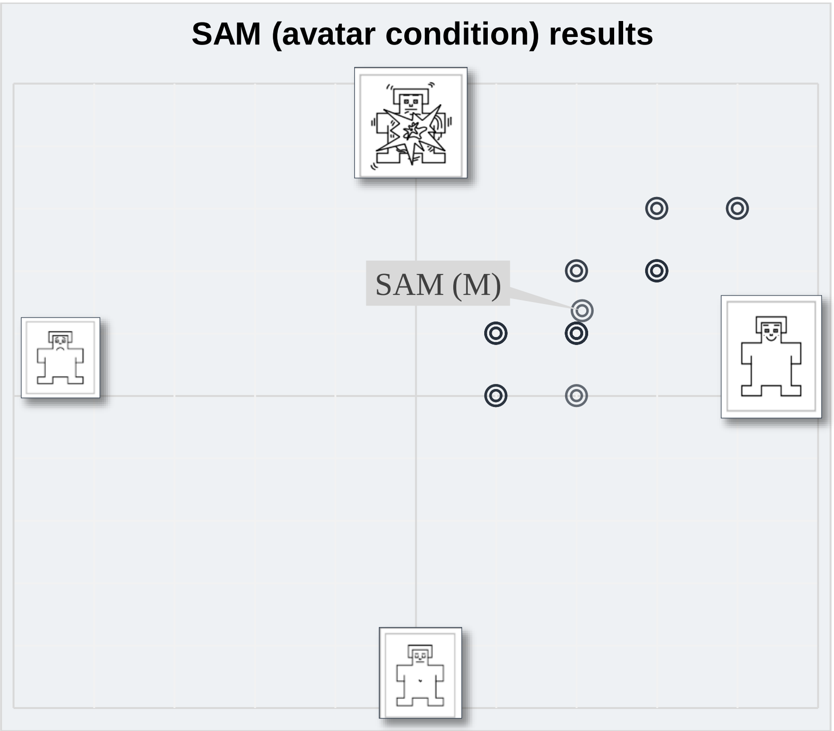


Figure 6-5: SAM result (avatar condition).

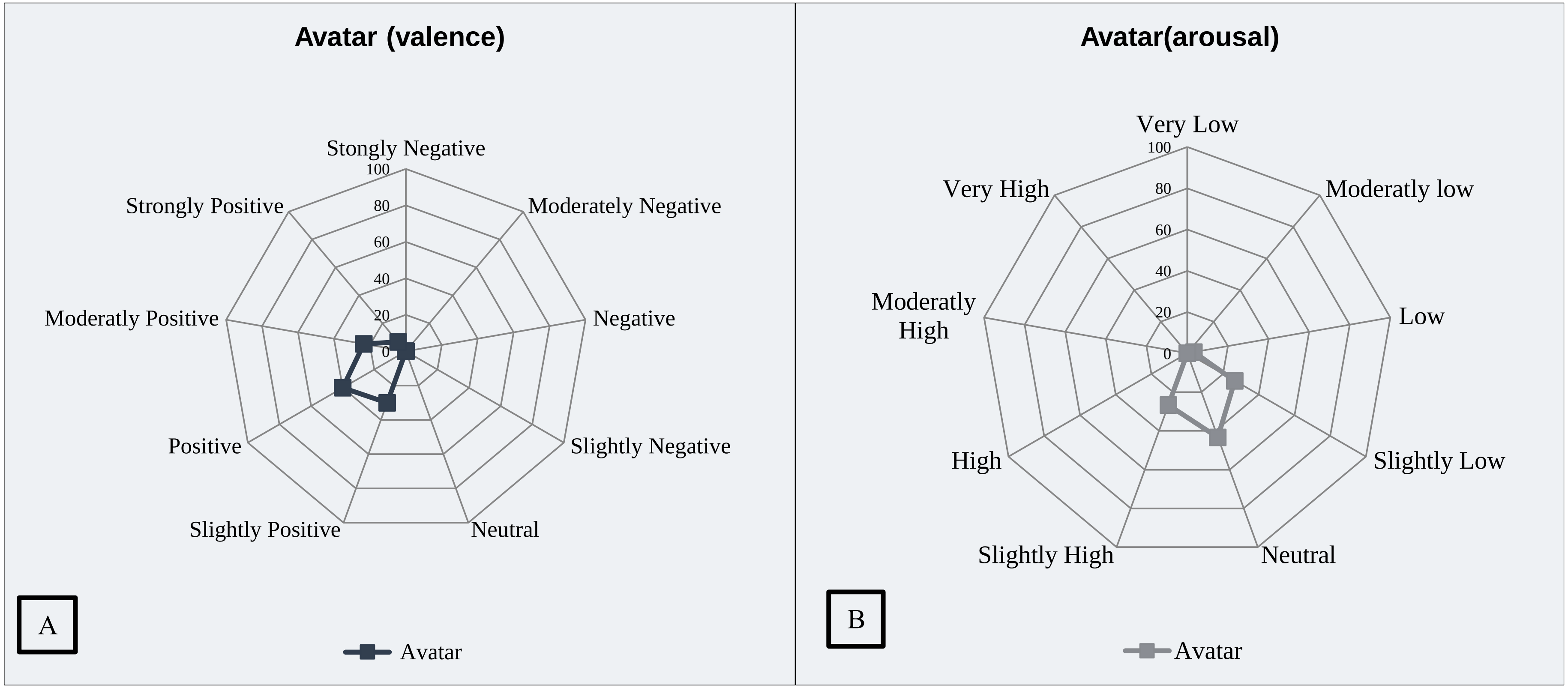


Figure 6-6: Avatar condition’s radar plots.

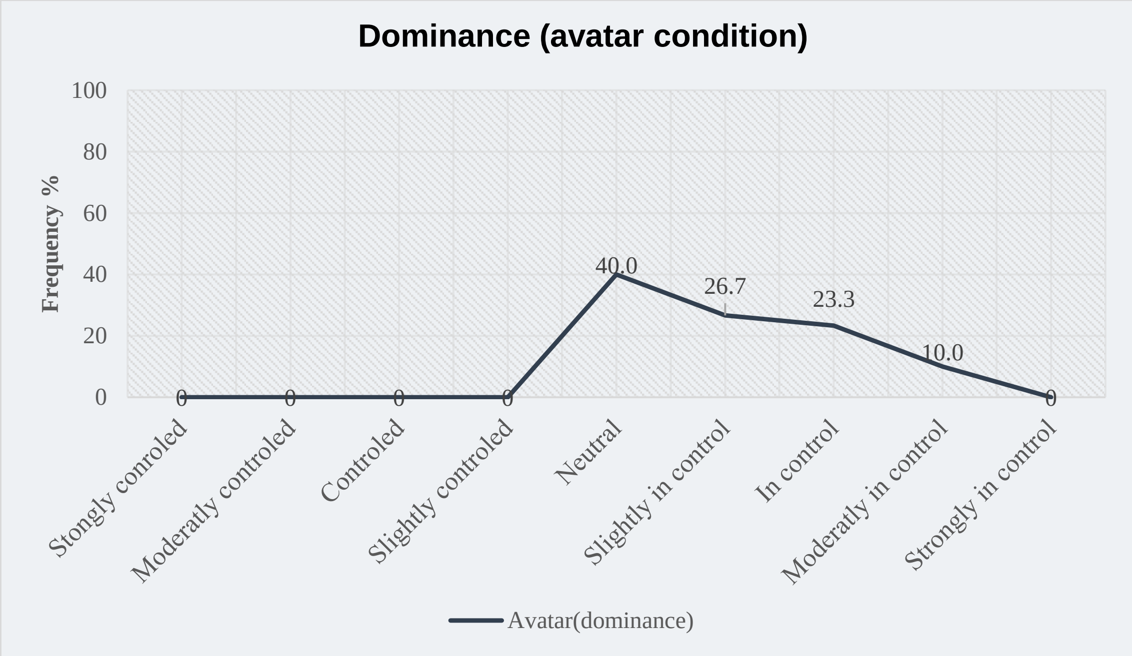


Figure 6-7: Dominance (avatar condition).

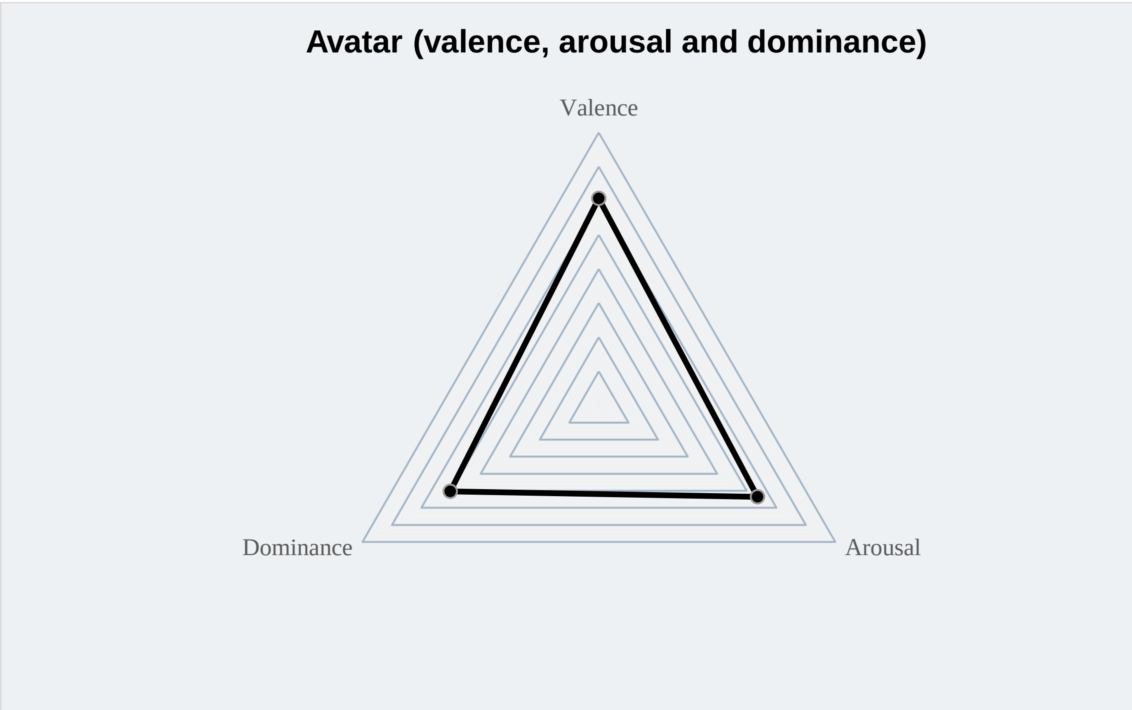


Figure 6-8: Avatar condition (valence, arousal and dominance).

The area covered with line graphs is small which indicates that there is considerable agreement among users in their ratings. Concerning the dominance dimension, (see Figure 6-7) user ratings were generally “neutral” with a tendency towards feeling “strongly in control”; the overall view of the SAM result is presented in Figure 6-8. A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users’ SAM dimensions. There was a strong positive correlation between the two dimensions (rs (28) = -.810, p <.001). The overall impression is that users felt the condition was positive and they were rather in control.

6.4.3 SAM (Earcons Condition)

The test results are displayed in Figure 6-9. Users rated the condition within the valence dimension as “negative” emotion towards the condition; within the arousal dimension users’ reaction is considered to be “neutral” and the the overall impression of the auditory condition was “negative”. For a better understanding of user ratings in both dimensions the results are plotted in radar charts (see Figure 6-10 (A,B)); user ratings within the valence dimension are moving from “negative” towards “slightly negtive”, while in the arousal dimension users ratings were placed at “slightly low“. The areas covered with line graphs is small which indicates that there is considerable agreement among users in their rating .

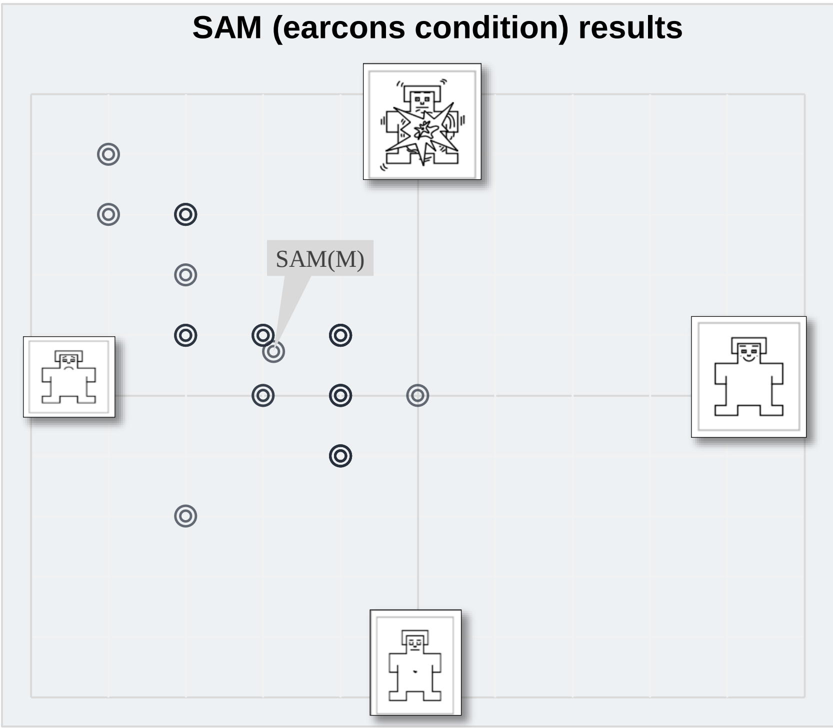


Figure 6-9: SAM result (earcons condition).

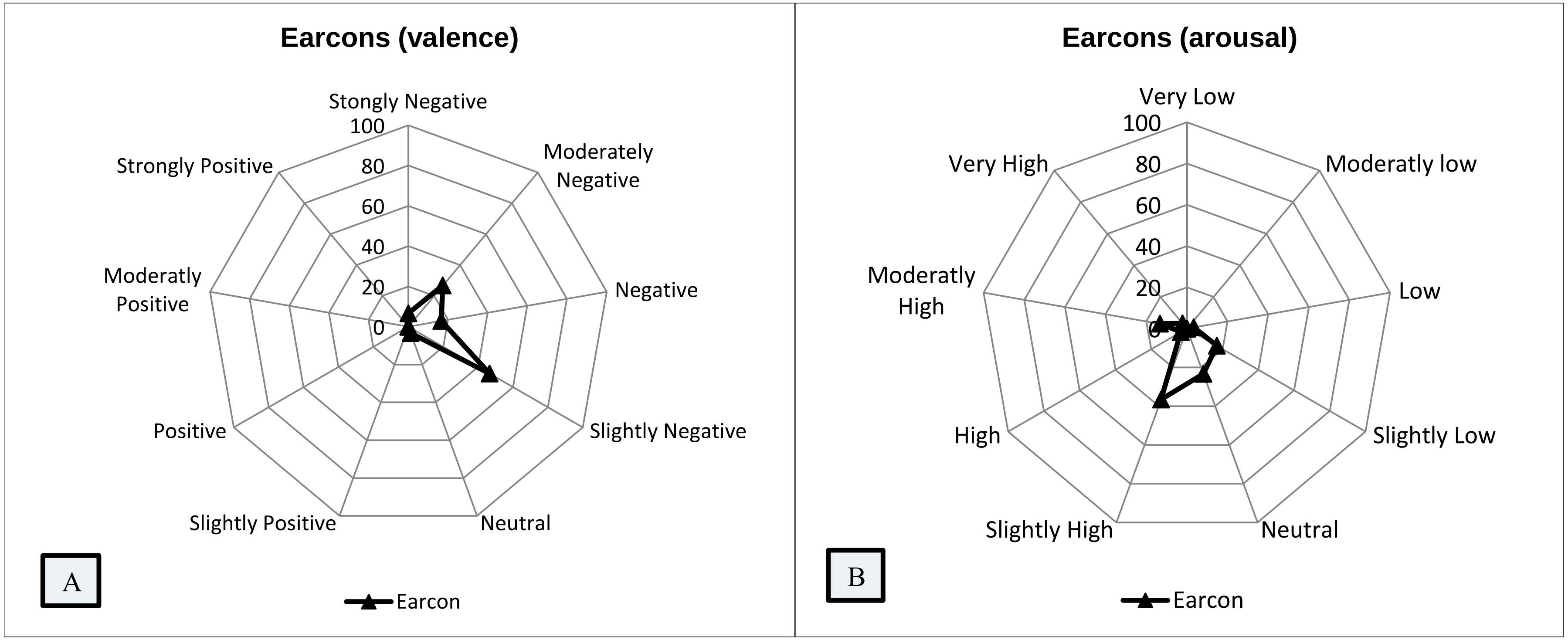


Figure 6-10: Earcons condition's radar plots.

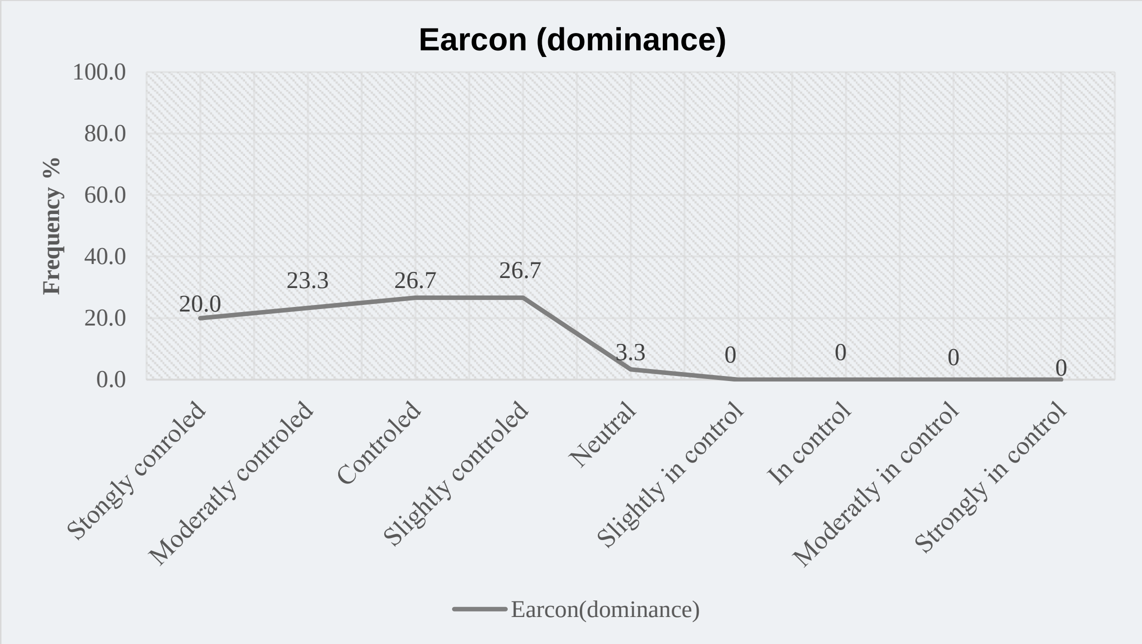


Figure 6-11: Dominance (earcons condition).

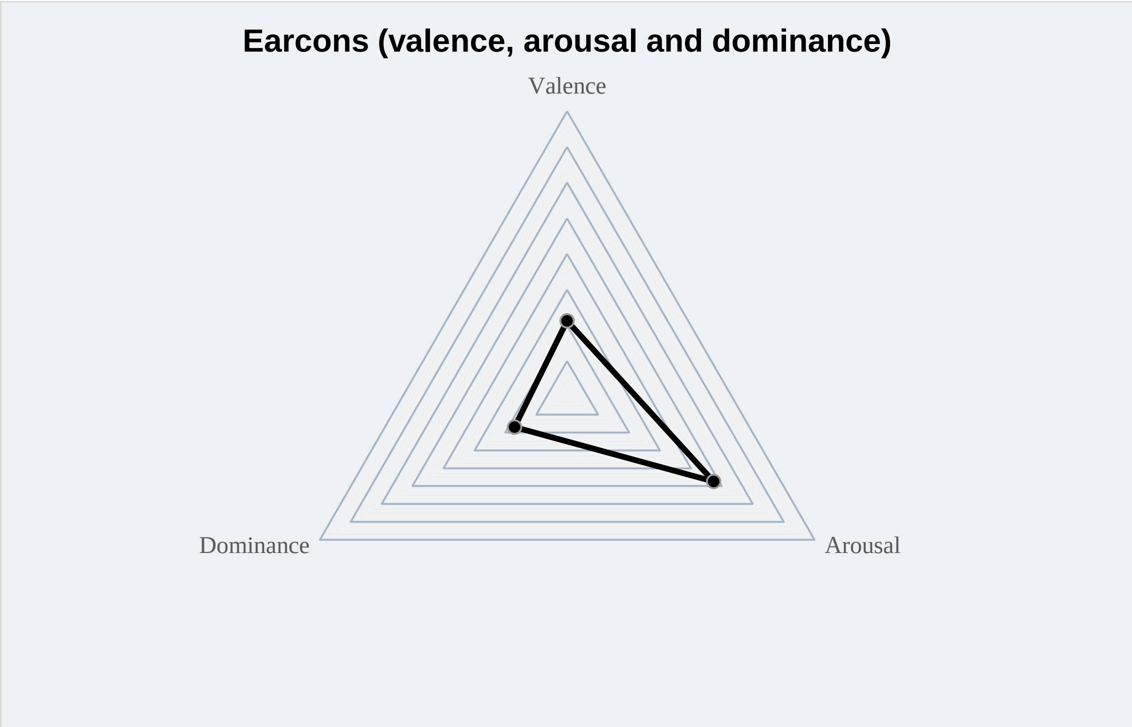


Figure 6-12: Earcons (valence, arousal and dominance).

Concerning the dominance dimension (see Figure 6.11), users ratings in general were "negative"; they felt they were rather "controlled." The overall view of SAM result is presented in Figure 6-12.

A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users' SAM dimensions. There was a strong negative correlation between the two dimensions (rs (28) = -.619, p <.001). The overall impression is that users felt the condition was negative and they were not in control.

6.4.4 SAM (Music Condition)

Users' ratings for the condition and the test results are displayed in Figure 6-13. Users regarded the condition within the valence dimension as "positive", whereas within the arousal dimension they considered it to be "neutral". The overall impression of the auditory condition was "positive". User ratings in both dimensions are plotted in radar charts (see Figure 6-14 (A,B)); the ratings for the valence dimension are moving from "neutral" towards "positive", while in the arousal dimension users' rating was placed between "low" and "high" with a tendency towards "neutral". The areas covered with line graphs is small which indicates that there is considerable agreement among users in their ratings.

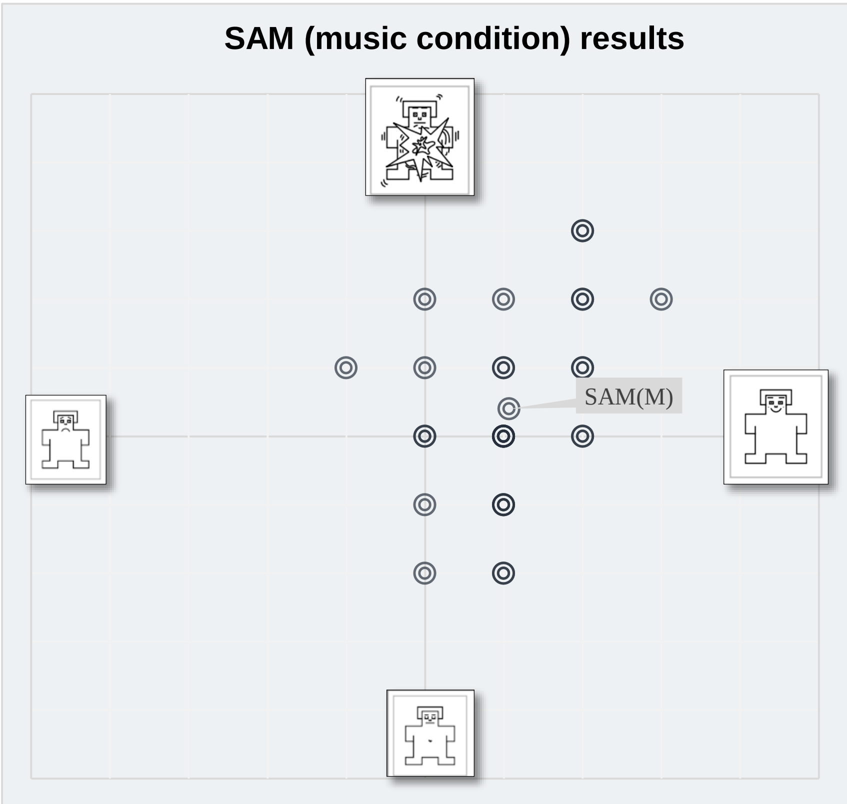


Figure 6-13: SAM result (music condition).

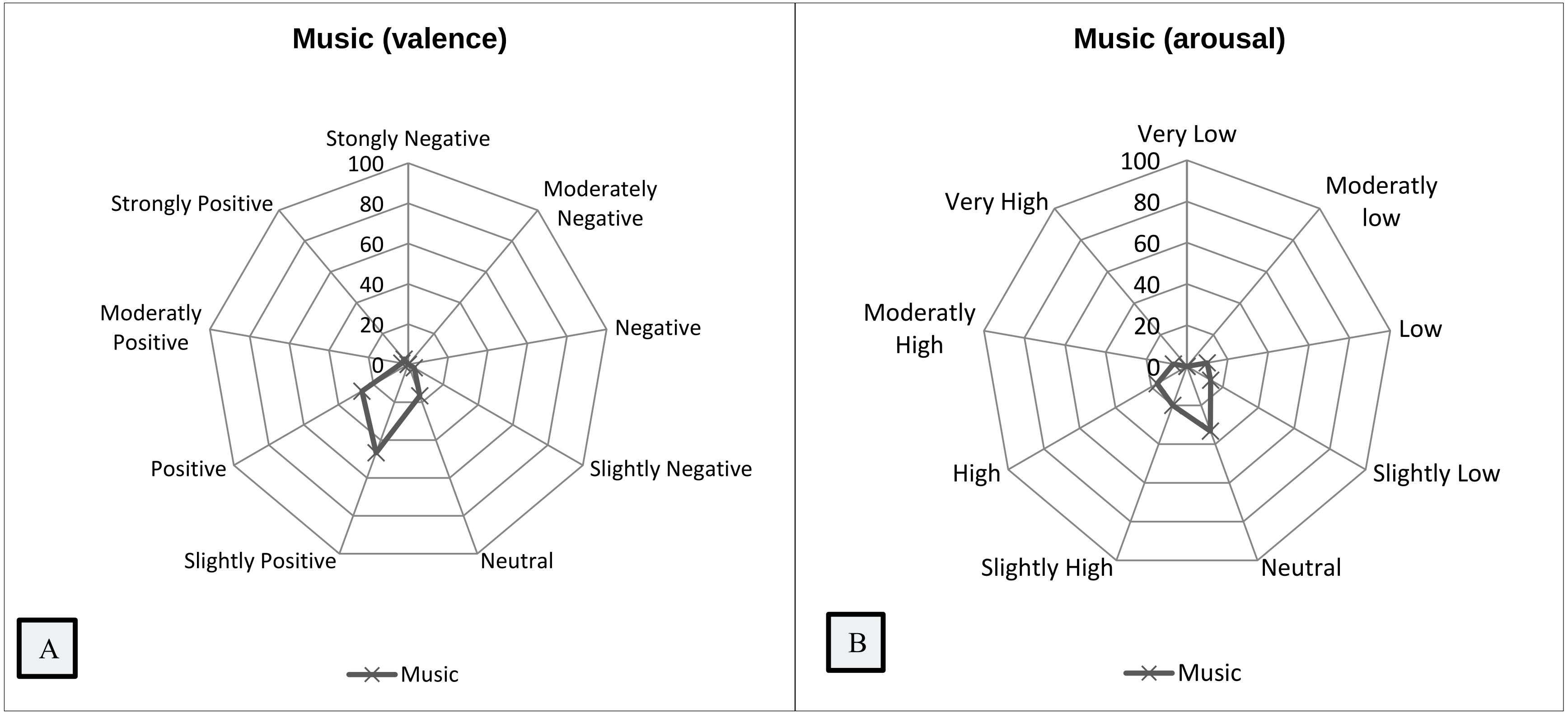


Figure 6-14: Music condition's radar plots.

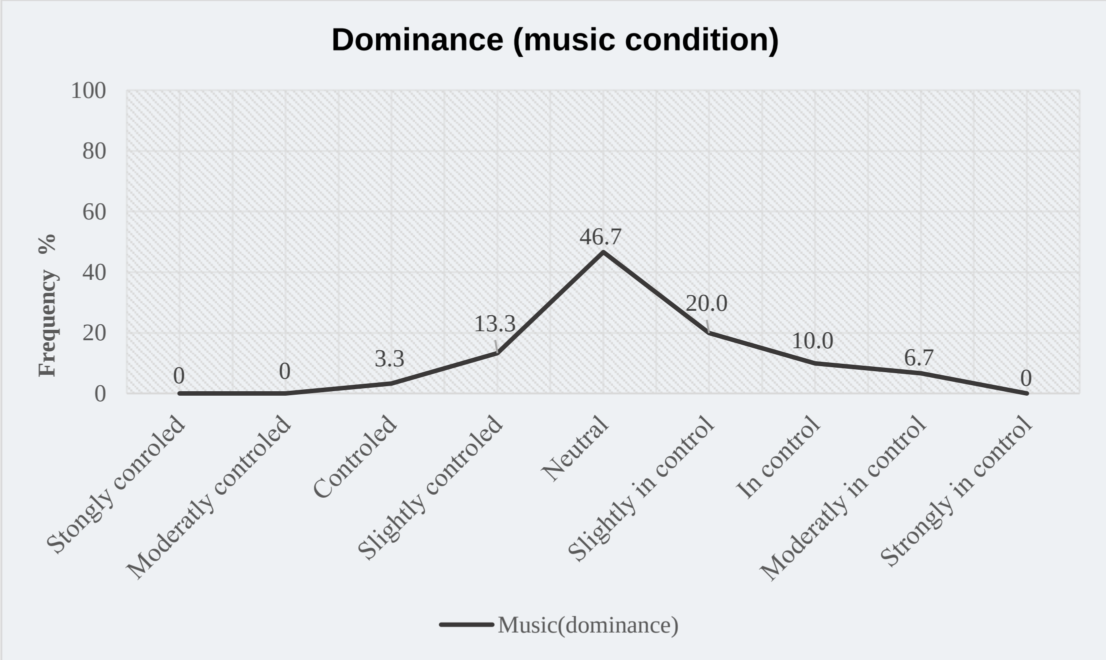


Figure 6-15: Dominance (music condition).

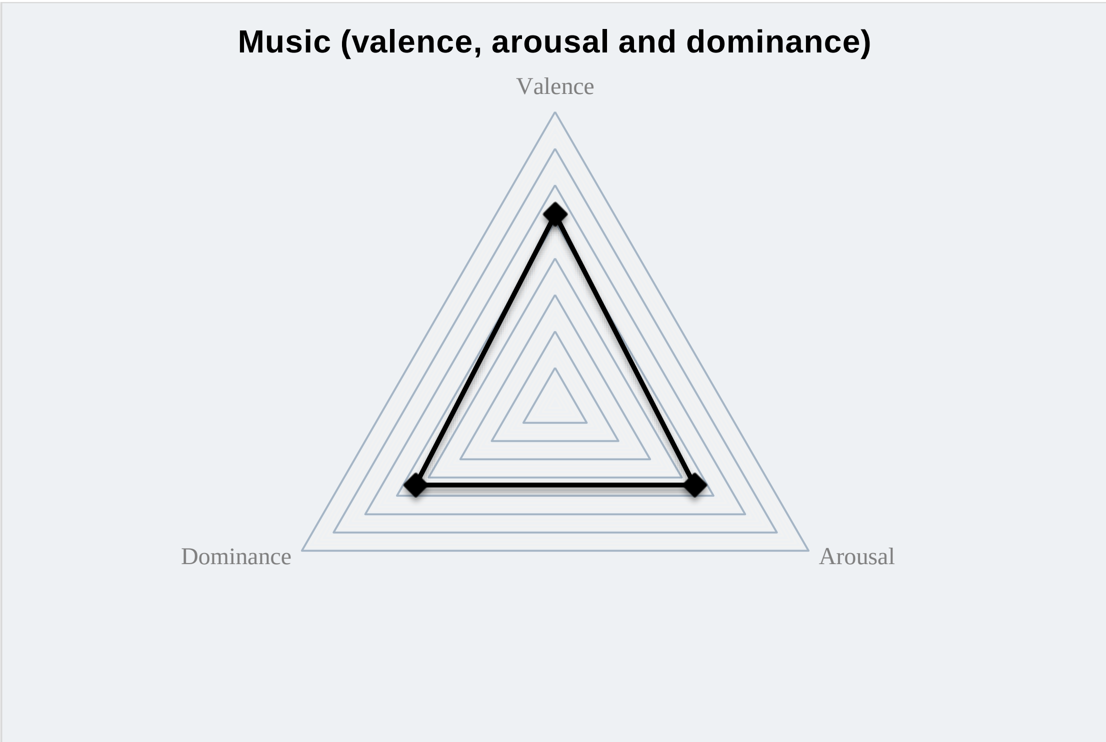


Figure 6-16: Music condition (valence, arousal and dominance)**.**

With reference to the dominance dimension (see Figure 6-15), user ratings were generally "neutral" with a tendency towards feeling "in control"; the overall view of SAM result is presented in Figure 6-16. A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users' SAM dimensions. There was strong positive correlation between the two dimensions (r_s (28) = .541, p <.005). The overall impression is that users felt the condition was positive and they were rather in control.

6.4.5 SAM (Speech Condition)

Users' ratings are displayed in Figure 6-17. Users' assesment of the condition within the valence dimension was "slightly positive" emotion towards the condition, within the arousal dimension users considered it to be "neutral" and the the overall impression of the auditory condition was "slightly positive". User ratings in both dimensions are plotted in radar charts (see Figure 6-18 (A,B)) user ratings for the valence dimension are moving from "neutral" towards "slightly positive", while in the arousal dimension user ratings were placed between "neutral" and "slightly high" with a tendency towards neutral. The areas covered with line graphs are small which indicates that there is considerable agreement among users in their rating .

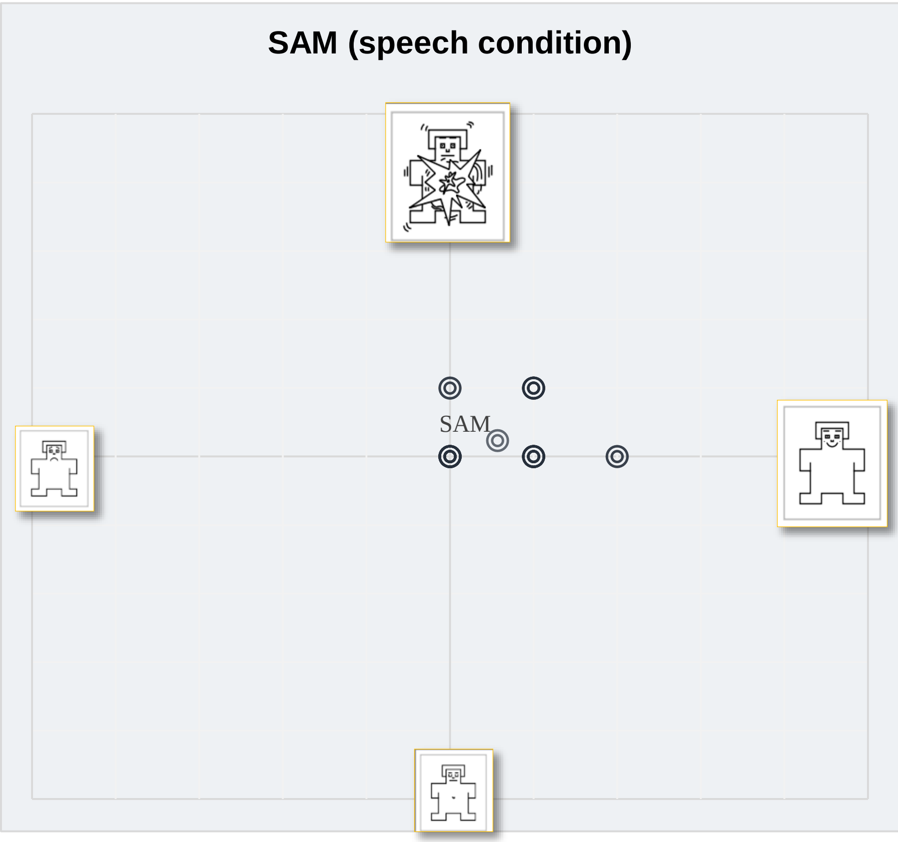


Figure 6-17: SAM result (speech condition).

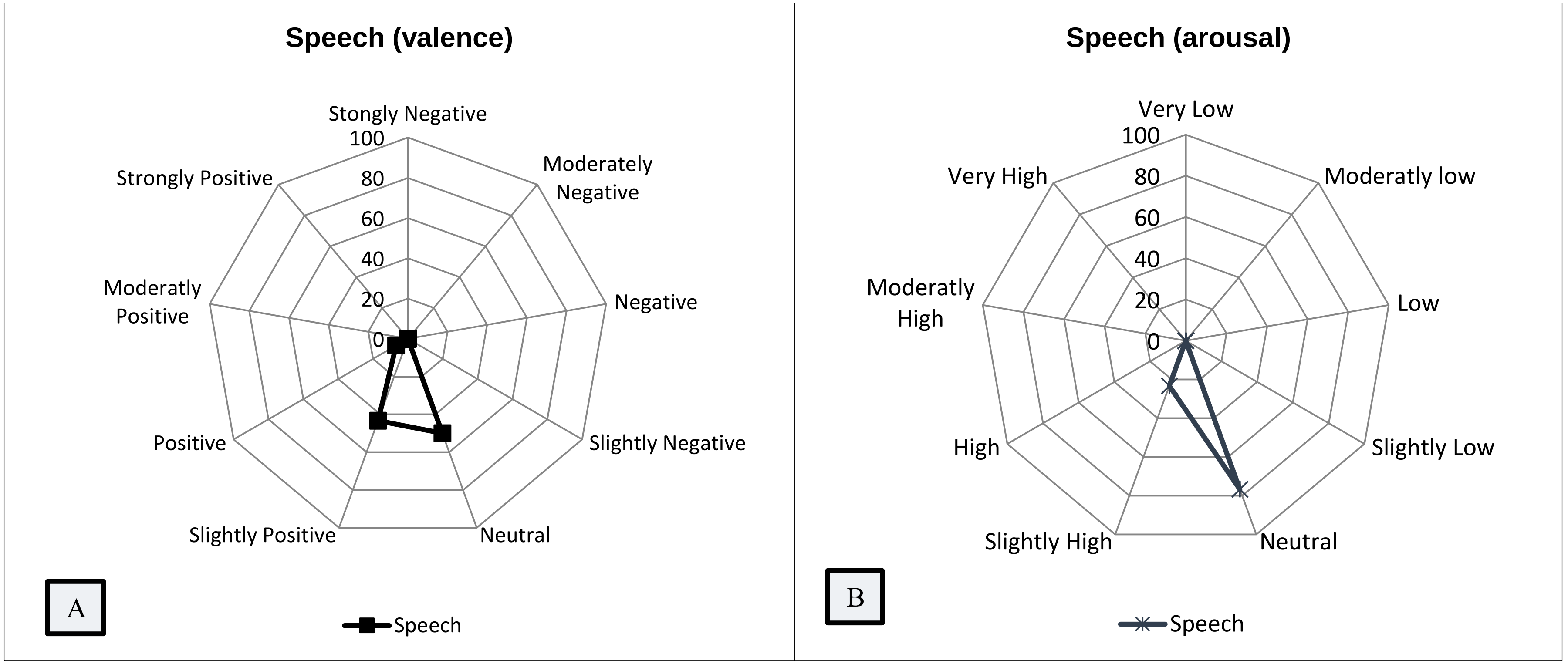


Figure 6-18: Speech condition’s radar plots.

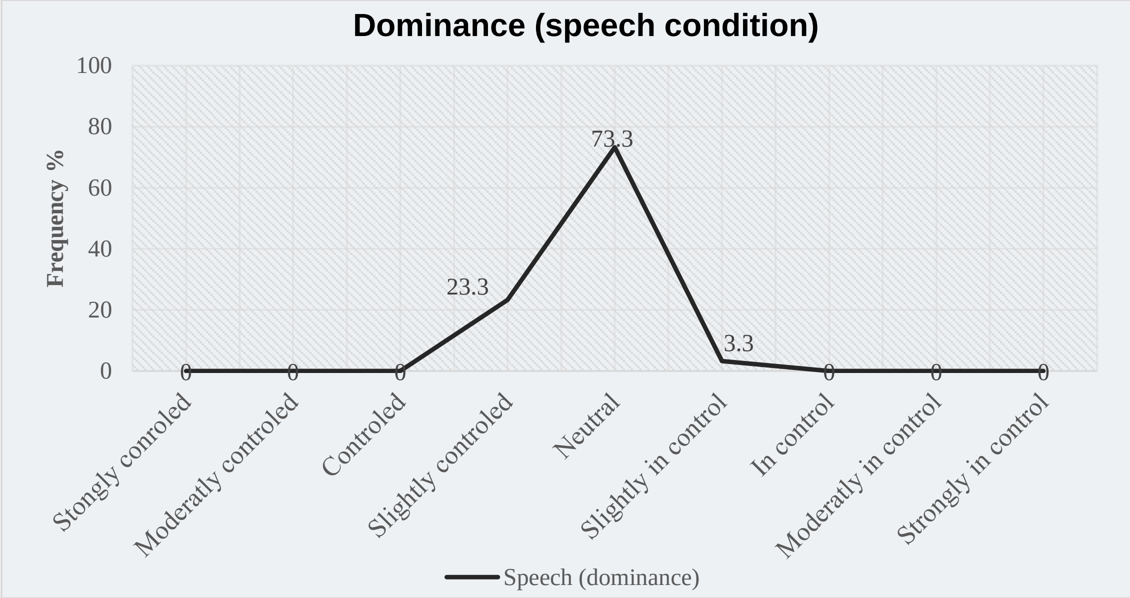


Figure 6-19: Dominance (speech condition).

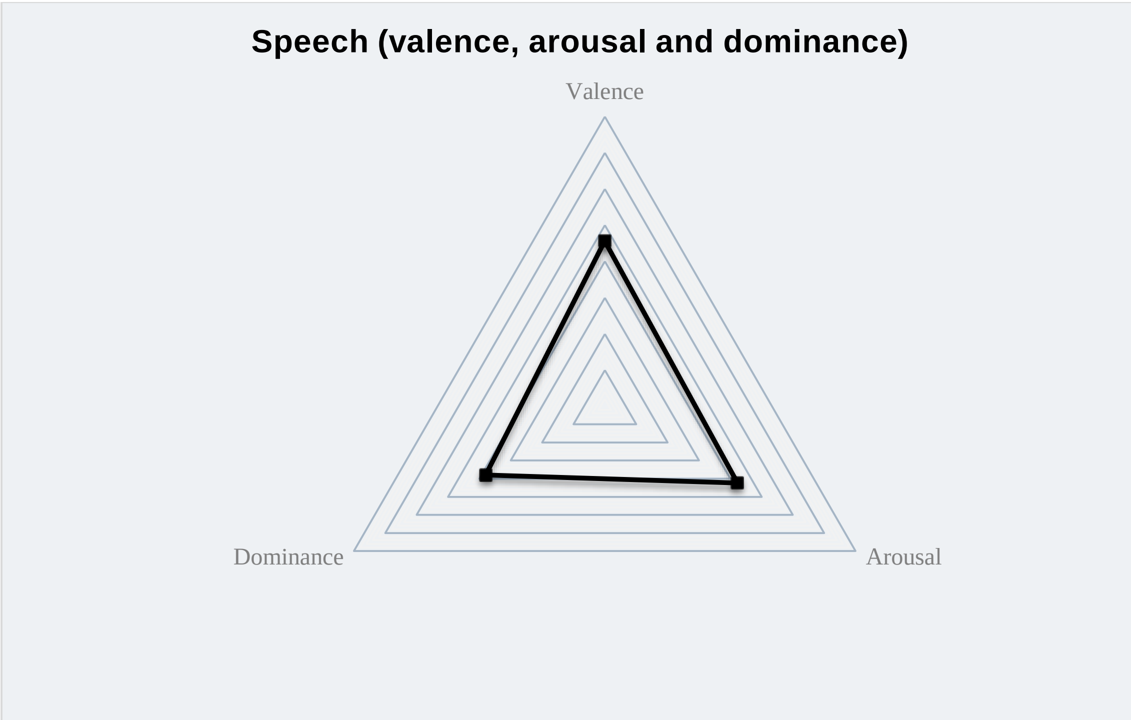


Figure 6-20: Speech (valence, arousal and dominance).

As for the dominance dimension (see Figure 6-19), user ratings were generally “neutral”; the overall view of the SAM result is presented in Figure 6-20. A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users’ SAM dimensions. There was a weak positive correlation between the two dimensions (rs (28)

= .179, p =.344). The overall impression is that users felt the condition was slightly positive and they were not completely in control.

6.4.6 SAM (NMMC)

Users' ratings for the condition were analysed and the test results are displayed in. Users' valuation of the condition within the valence dimension was "slightly negative" emotion towards the condition, within the arousal dimension users reaction is considered to be "slightly low" and the the overall impression of the auditory condition was "slightly negative". For a better understanding of user ratings in both dimensions the results are plotted in radar charts (see Figure 6-22 (A and B)); users ratings' for the valence dimension are moving from "negative" towards "slightly negative" as well as "neutral", while within the arousal dimension users ratings were placed between "low" and "neutral" with a tendency towards "slightly high".

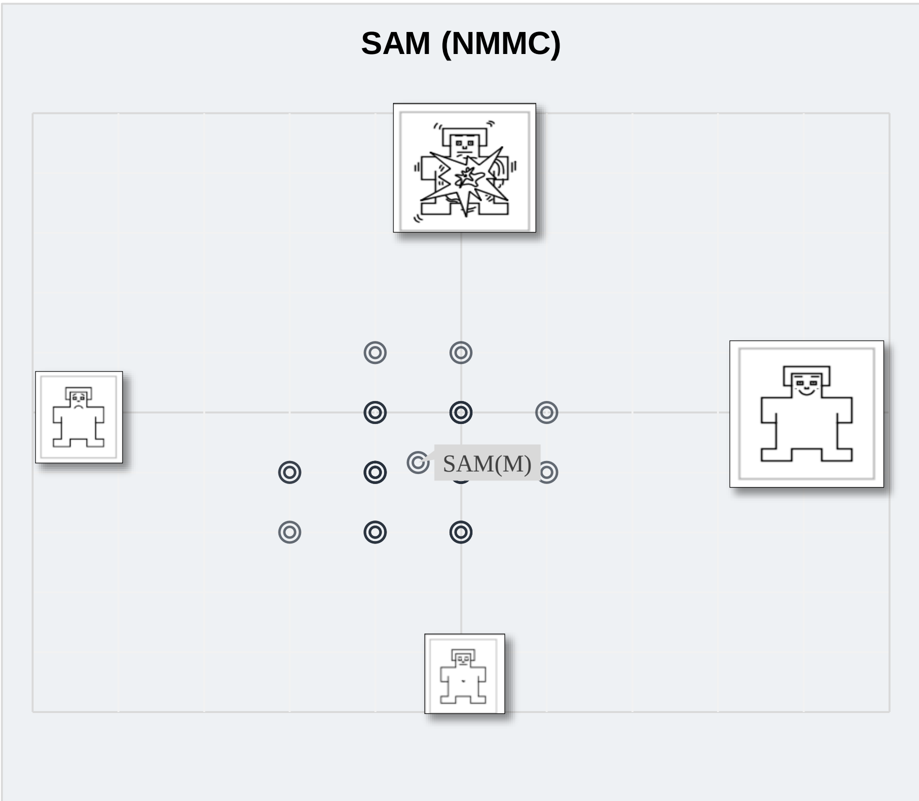
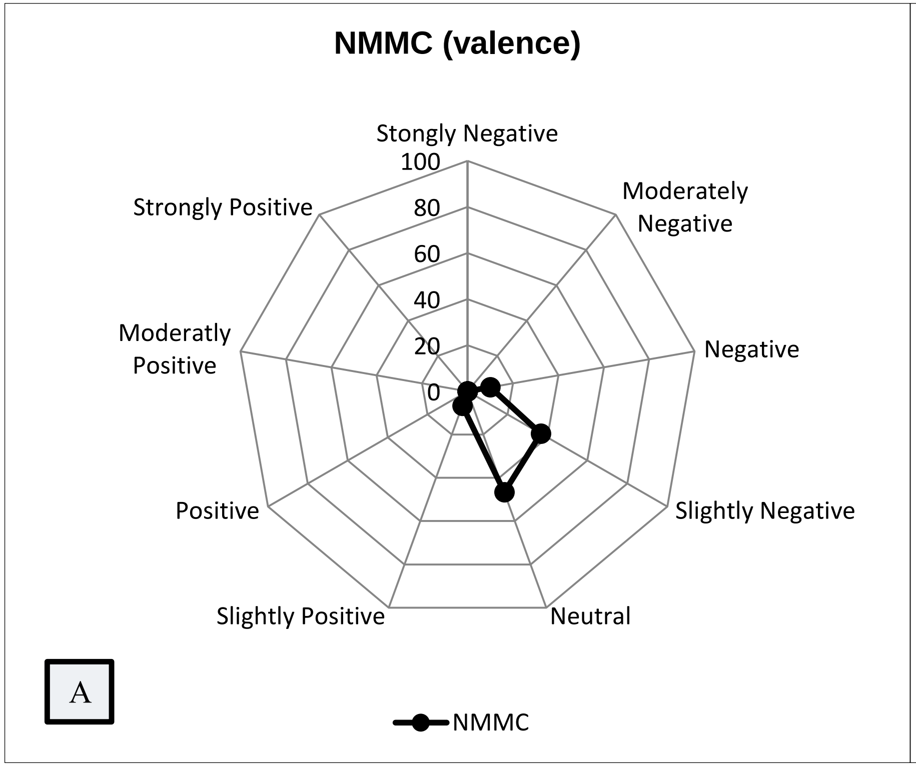


Figure 6-21: SAM result (NMMC).



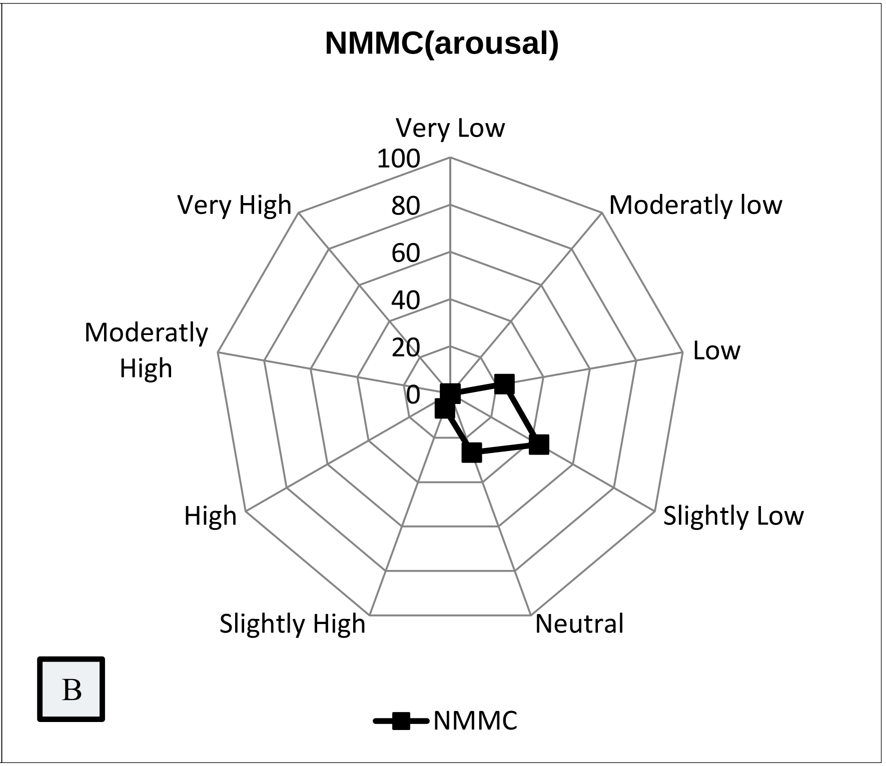


Figure 6-22: NMMC radar plots.

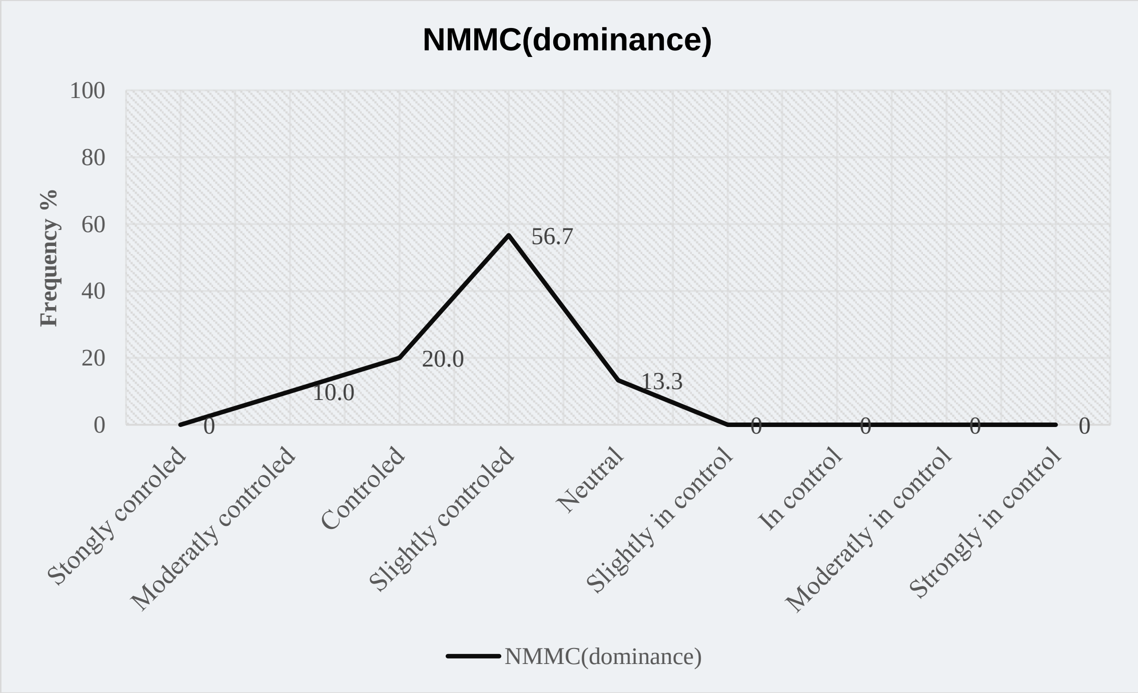


Figure 6-23: Dominance (NMMC) results.

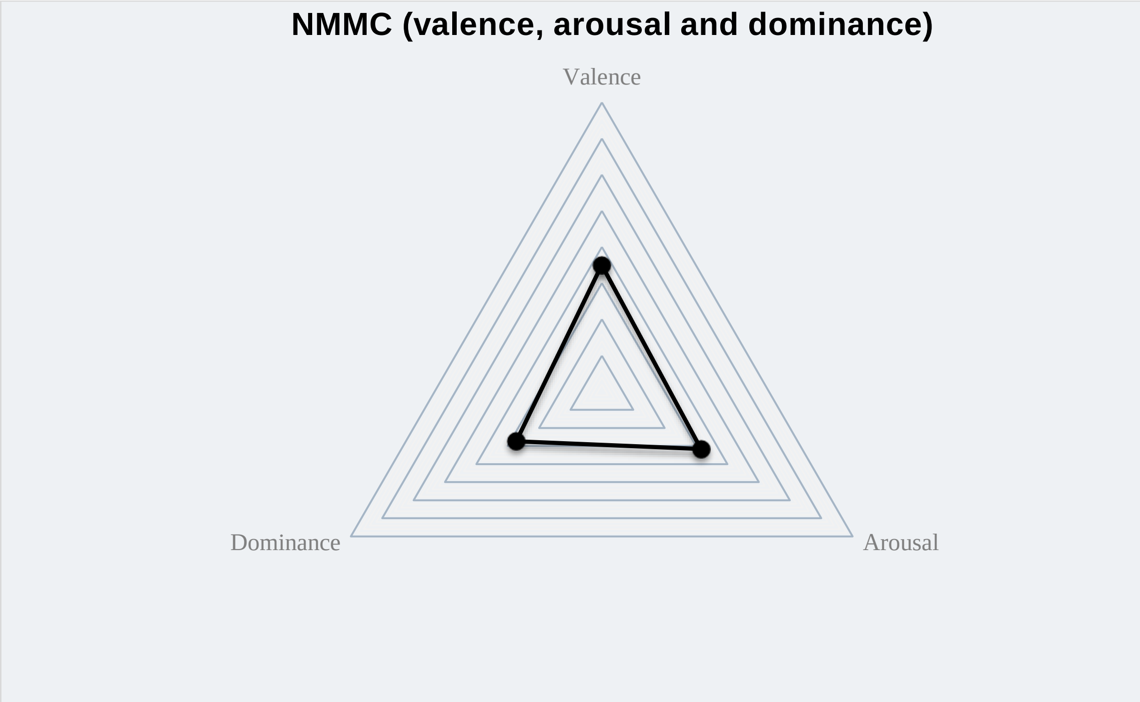


Figure 6-24: NMMC (valence, arousal and dominance).

The areas covered with line graphs are small which indicates that there is considerable agreement among users in their rating. Concerning the dominance dimension (see Figure 6-23), user ratings were generally “slightly negative” with a tendency towards feeling "controlled"; the overall view of the SAM result is presented in Figure 6-24. A Spearman's rank-order correlation was run to assess the relationship between arousal and valence for users' SAM dimensions. There was a weak positive correlation between the two dimensions (r_s (28) = .172, p =.362). Overall, users felt the condition was slightly negative and that they were not completely in control.

6.4.7 SAM (All Conditions)

Figure 6-25 presents an overall comparison between users' ratings in the three dimensions of SAM (valence, arousal and dominance). Within the valence dimension, the most positive condition was the avatar condition (M=7.06) followed by the music condition (M=6.2) then the speech condition (M=5.67) and these were more positive than the control condition (NMMC, M=4.5).

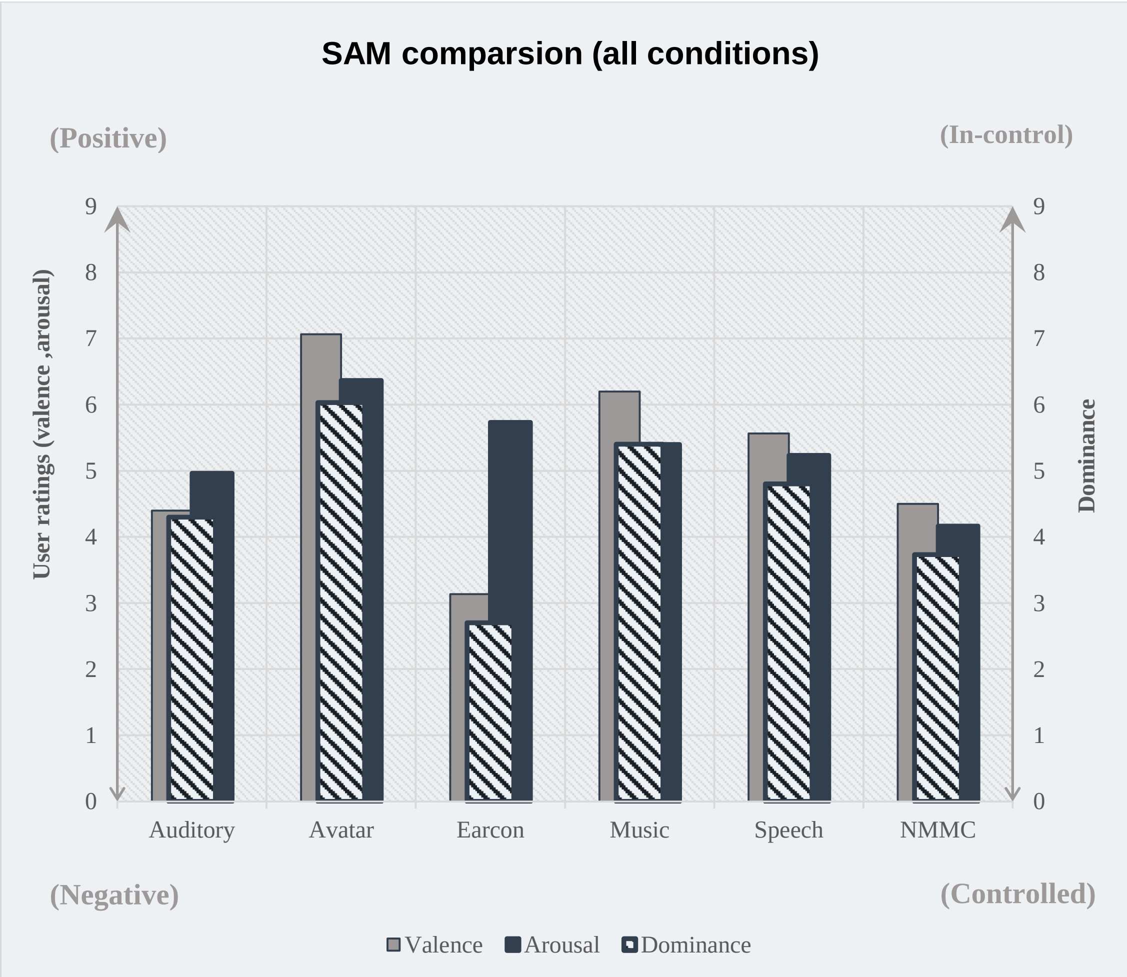


Figure 6-25: SAM's' comparison (all conditions).

While the remaining conditions were less positive than the control condition, the earcons condition was the least positive (M=3.13) followed by the auditory condition (M=4.4). Concerning the arousal dimension, it is noted that all of the experimented conditions were more stimulating than the control condition (M=4.16). The avatar condition was rated as the highest (M=6.37) followed by the earcons condition (M=5.73), the music condition (M=5.40), then the speech condition (M=5.23) and the auditory condition (M=4.97).

6.4.8 **Statistical Data Analysis**

The aim of the statistical analysis is to examine the data and find out if the perceived differences are statically significant and are not due to chance. The results in the SAM dimensions for all the conditions will be examined.

6.4.8.1 ***Valence***

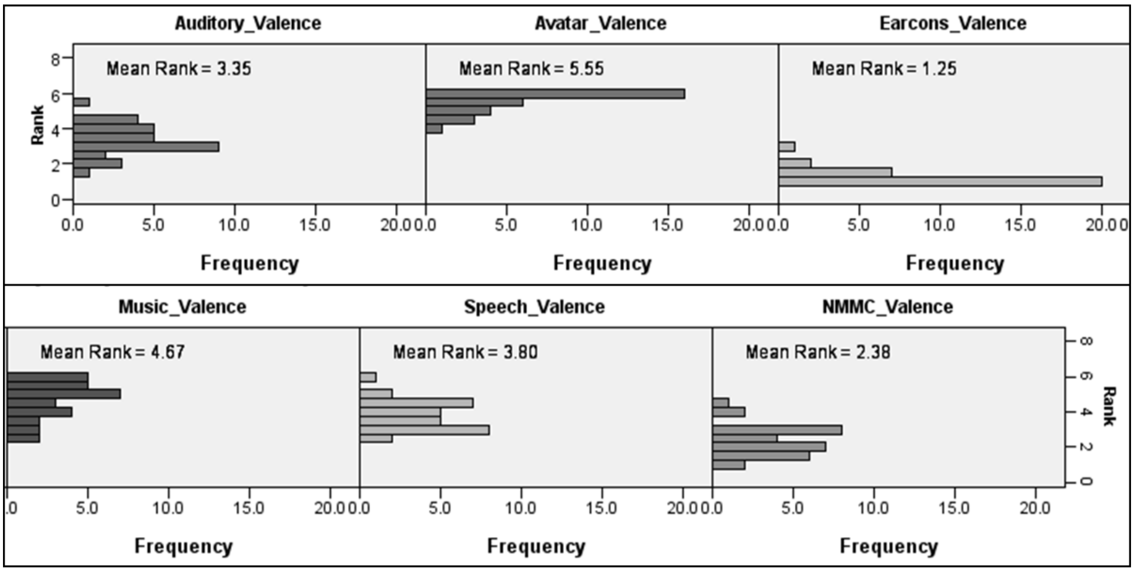
To decide on the appropriate statistical test the data had to be checked to see if it is normally distributed. The “Shapiro-Wilk” normality test was used and the results shown in Table 6-1 indicate that the data did not fit the normality criteria.

Tests of normality						
Valence	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.253	30	.000	.796	30	.0001
Avatar	.229	30	.000	.860	30	.001
Earcons	.290	30	.000	.844	30	.0001
NMMC	.273	30	.000	.853	30	.001
Music	.246	30	.000	.897	30	.007
Speech	.317	30	.000	.742	30	.0001
a. Lilliefors Significance Correction						

Table 6-1: Valence tests of normality (all conditions).

Considering the normality test results, in addition to the number of conditions and the fact that the study uses repeated measurements, the appropriate statistical test is Friedmans’ ANOVA test (see the test result in Figure 6-26). A *post-hoc* test was needed to identify where the differences occurred; therefore, pairwise comparisons were conducted and the results are presented in Figure 6-28 (significant differences are denoted in gray). For a better understanding of the test results, the results of the pairwise comparisons are presented in Figure 6-28. They are shown in ascending order according to their median and mean ranks.

Friedmans' ANOVA test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory_Valence, Avatar_Valence, Earcons_Valence, Music_Valence, Speech_Valence and NMMC_Valence are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	113.667
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 6-26: Friedmans' ANOVA test of valence (all conditions).

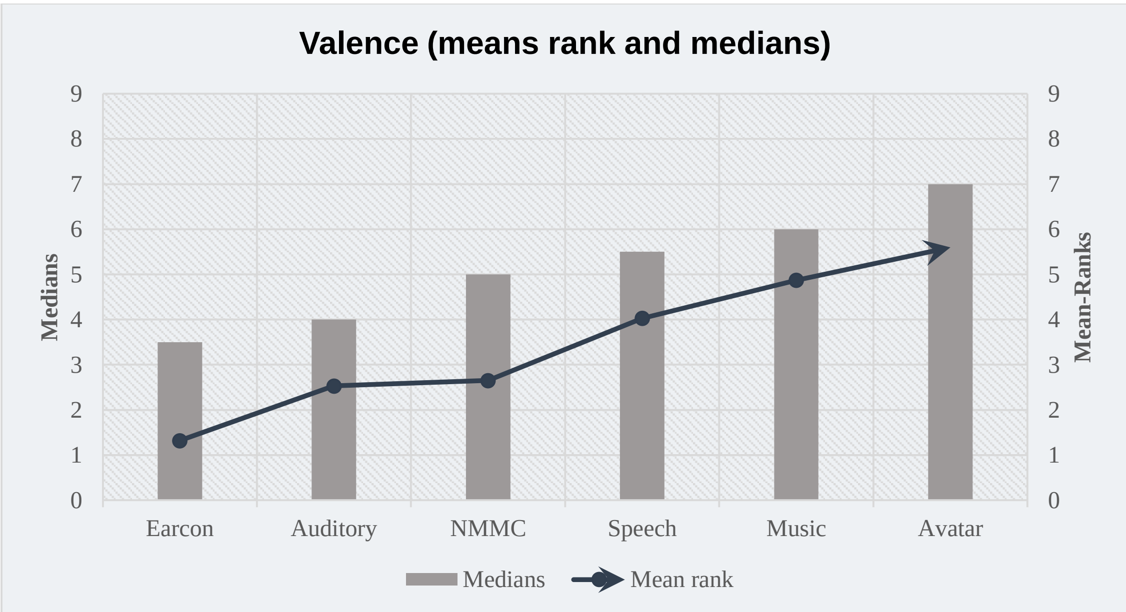
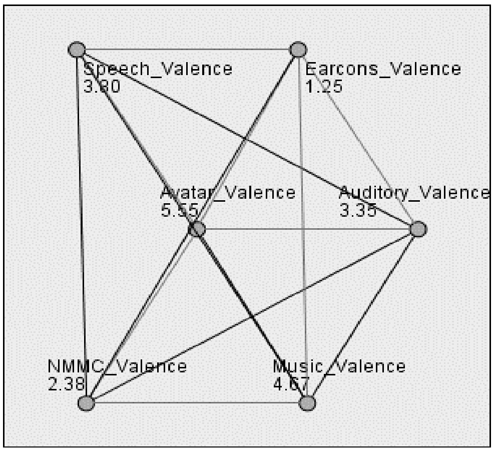


Figure 6-27: Valence means ranks and medians.

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Earcons_Valence-NMMC_Valence	-1.133	.483	-2.346	.019	.284
Earcons_Valence-Auditory_Valence	2.100	.483	4.347	.000	.000
Earcons_Valence-Speech_Valence	-2.550	.483	-5.279	.000	.000
Earcons_Valence-Music_Valence	-3.417	.483	-7.073	.000	.000
Earcons_Valence-Avatar_Valence	4.300	.483	8.902	.000	.000
NMMC_Valence-Auditory_Valence	.967	.483	2.001	.045	.681
NMMC_Valence-Speech_Valence	1.417	.483	2.933	.003	.050
NMMC_Valence-Music_Valence	2.283	.483	4.727	.000	.000
NMMC_Valence-Avatar_Valence	3.167	.483	6.556	.000	.000
Auditory_Valence-Speech_Valence	-.450	.483	-.932	.352	1.000
Auditory_Valence-Music_Valence	-1.317	.483	-2.726	.006	.096
Auditory_Valence-Avatar_Valence	-2.200	.483	-4.554	.000	.000
Speech_Valence-Music_Valence	.867	.483	1.794	.073	1.000
Speech_Valence-Avatar_Valence	1.750	.483	3.623	.000	.004
Music_Valence-Avatar_Valence	.883	.483	1.829	.067	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6-28: Valence pairwise comparisons.

6.4.8.2 *Arousal*

To select the appropriate statistical test, “Shapiro-Wilk” normality test was used and the results shown in Table 6-2 indicated that the data did not fit the normality criteria. Considering this, in addition to the number of conditions and the fact that the study uses repeated measurements, the appropriate statistical test is Freedman’s ANOVA test. The test results are presented in Figure 6-29.

Tests of normality						
Condition Arousal	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.232	30	.001	.857	30	.001
Avatar	.293	30	.001	.856	30	.001
Earcons	.226	30	.001	.924	30	.034
Music	.181	30	.014	.939	30	.088
Speech	.473	30	.001	.526	30	.001
NMMC	.242	30	.001	.870	30	.002
a. Lilliefors Significance Correction						

Table 6-2: Normality test (arousal).

Friedmans' ANOVA test

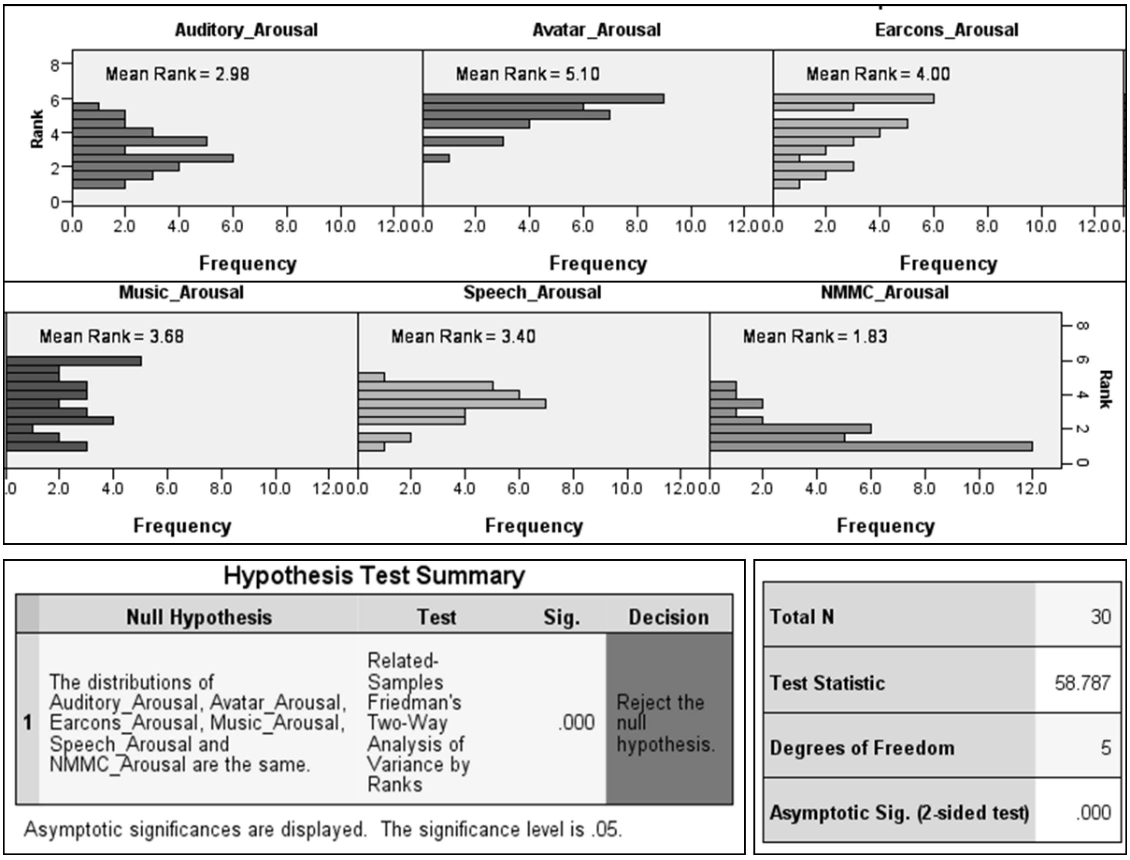
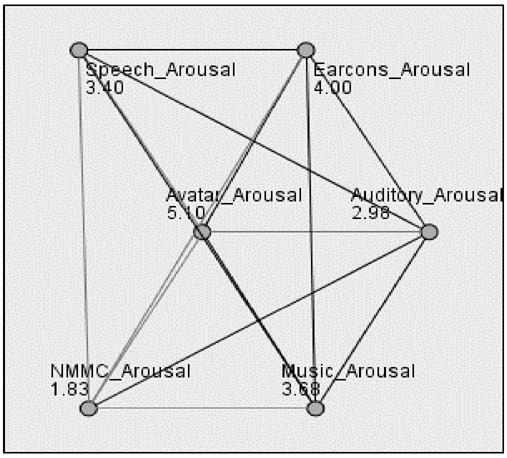


Figure 6-29: Friedmans' ANOVA test of arousal (all conditions).

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
NMMC_Arousal-Auditory_Arousal	1.150	.483	2.381	.017	.259
NMMC_Arousal-Speech_Arousal	1.567	.483	3.243	.001	.018
NMMC_Arousal-Music_Arousal	1.850	.483	3.830	.000	.002
NMMC_Arousal-Earcons_Arousal	2.167	.483	4.485	.000	.000
NMMC_Arousal-Avatar_Arousal	3.267	.483	6.763	.000	.000
Auditory_Arousal-Speech_Arousal	-.417	.483	-.863	.388	1.000
Auditory_Arousal-Music_Arousal	-.700	.483	-1.449	.147	1.000
Auditory_Arousal-Earcons_Arousal	-1.017	.483	-2.105	.035	.530
Auditory_Arousal-Avatar_Arousal	-2.117	.483	-4.382	.000	.000
Speech_Arousal-Music_Arousal	.283	.483	.587	.558	1.000
Speech_Arousal-Earcons_Arousal	.600	.483	1.242	.214	1.000
Speech_Arousal-Avatar_Arousal	1.700	.483	3.519	.000	.006
Music_Arousal-Earcons_Arousal	.317	.483	.656	.512	1.000
Music_Arousal-Avatar_Arousal	1.417	.483	2.933	.003	.050
Earcons_Arousal-Avatar_Arousal	1.100	.483	2.277	.023	.342

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6-30: Arousal pairwise comparisons.

A *post-hoc* test was needed to identify where the differences occurred therefore pairwise comparisons were conducted. The results are presented in Figure 6-30 (significant differences are denoted in gray. For a better understanding of the test results, the pairwise comparisons results are presented in Figure 6-31 they are presented in ascending order according to their median and mean ranks.

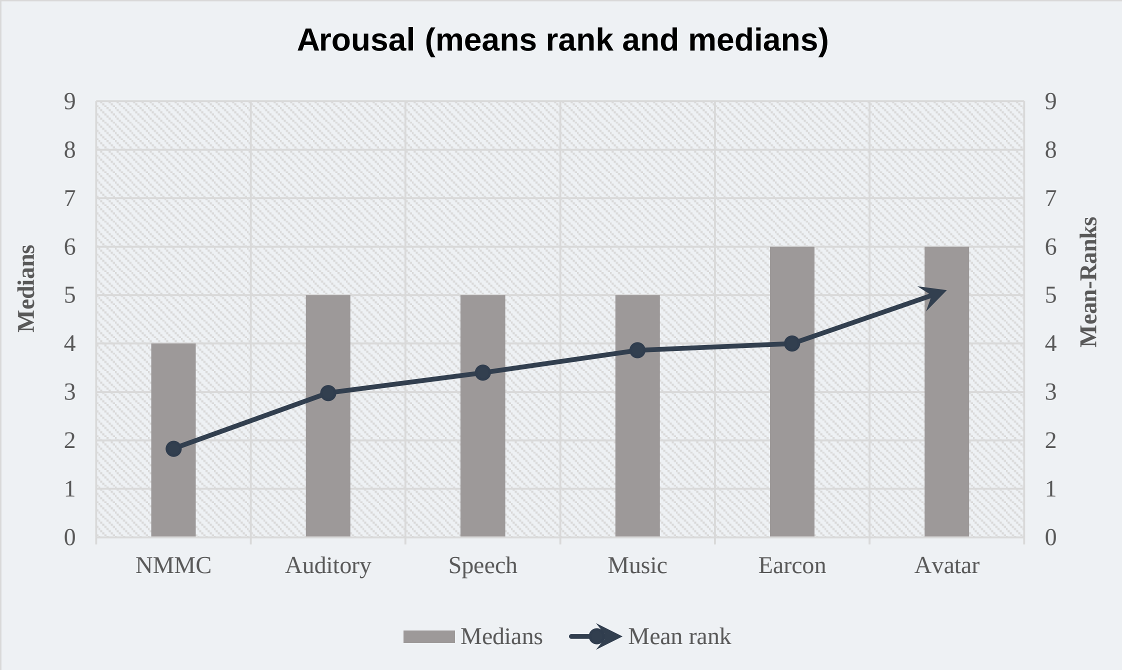


Figure 6-31: Arousal means rank and medians.

6.4.8.3 ***Dominance***

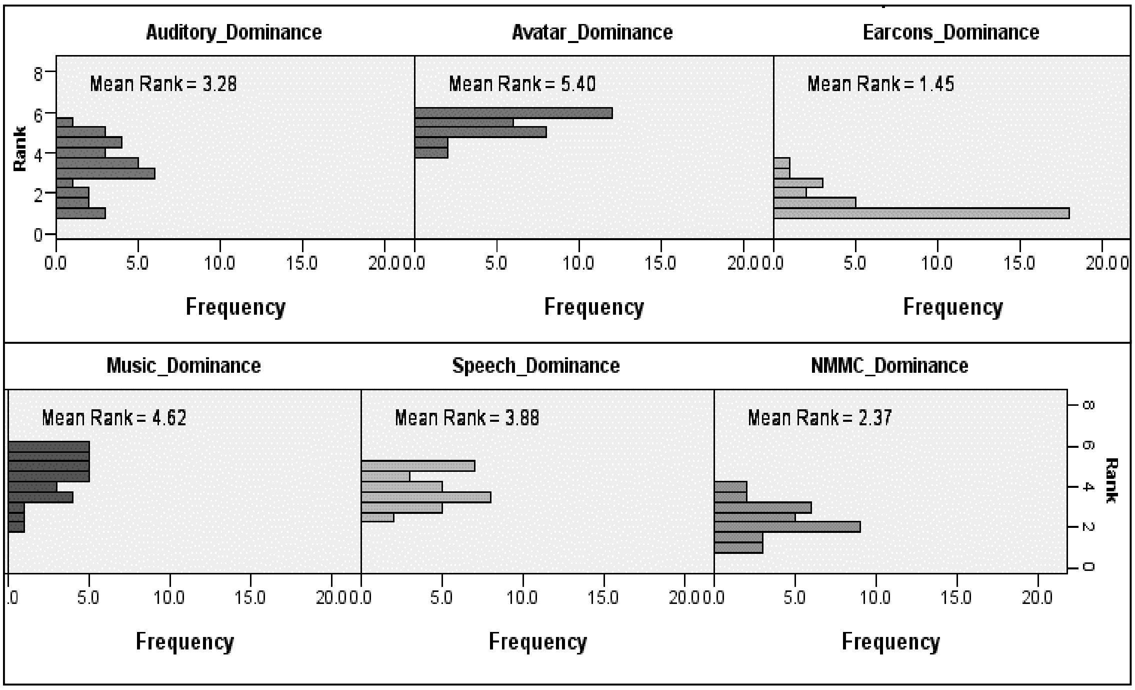
To decide on the appropriate statistical test the data had to be checked to see if it was normally distributed. The Shapiro-Wilk normality test was used and the results shown in Table 6-3 indicate that the data did not fit the normality criteria.

Tests of Normality						
Condition	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
Dominance	Statistic	df	Sig.	Statistic	df	Sig.
Auditory	.313	30	.000	.815	30	.000
Avatar	.241	30	.000	.834	30	.000
Earcons	.167	30	.032	.899	30	.008
Music	.268	30	.000	.901	30	.009
Speech	.427	30	.000	.646	30	.000
NMMC	.326	30	.000	.829	30	.000
a. Lilliefors Significance Correction						

Table 6-3: Normality test (dominance).

Considering this, in addition to the number of conditions and the fact that the study uses repeated measurements, the appropriate statistical test is Freedman's ANOVA test. The test results are presented in Figure 6-32.

Friedmans' ANOVA test



Hypothesis Test Summary

	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory_Dominance, Avatar_Dominance, Earcons_Dominance, Music_Dominance, Speech_Dominance and NMMC_Dominance are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

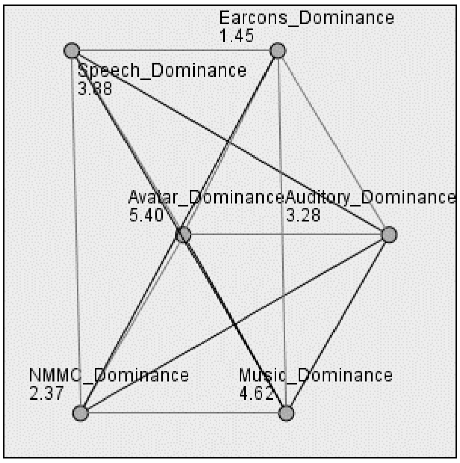
Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	102.641
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 6-32: Friedmans' ANOVA test of dominance (all conditions).

A *post-hoc* test was needed to identify where the differences occurred therefore pairwise comparisons were conducted and the results are presented in Figure 6-33 (significant differences are denoted in gray). For better understanding of the differences among the tested conditions see Figure 6-34 where they are ranked according to their means and medians.

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Earcons_Dominance-NMMC_Dominance	-.917	.483	-1.898	.058	.866
Earcons_Dominance-Auditory_Dominance	1.833	.483	3.795	.000	.002
Earcons_Dominance-Speech_Dominance	-2.433	.483	-5.037	.000	.000
Earcons_Dominance-Music_Dominance	-3.167	.483	-6.556	.000	.000
Earcons_Dominance-Avatar_Dominance	3.950	.483	8.177	.000	.000
NMMC_Dominance-Auditory_Dominance	.917	.483	1.898	.058	.866
NMMC_Dominance-Speech_Dominance	1.517	.483	3.140	.002	.025
NMMC_Dominance-Music_Dominance	2.250	.483	4.658	.000	.000
NMMC_Dominance-Avatar_Dominance	3.033	.483	6.280	.000	.000
Auditory_Dominance-Speech_Dominance	-.600	.483	-1.242	.214	1.000
Auditory_Dominance-Music_Dominance	-1.333	.483	-2.760	.006	.087
Auditory_Dominance-Avatar_Dominance	-2.117	.483	-4.382	.000	.000
Speech_Dominance-Music_Dominance	.733	.483	1.518	.129	1.000
Speech_Dominance-Avatar_Dominance	1.517	.483	3.140	.002	.025
Music_Dominance-Avatar_Dominance	.783	.483	1.622	.105	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6-33: Dominance pairwise comparisons.

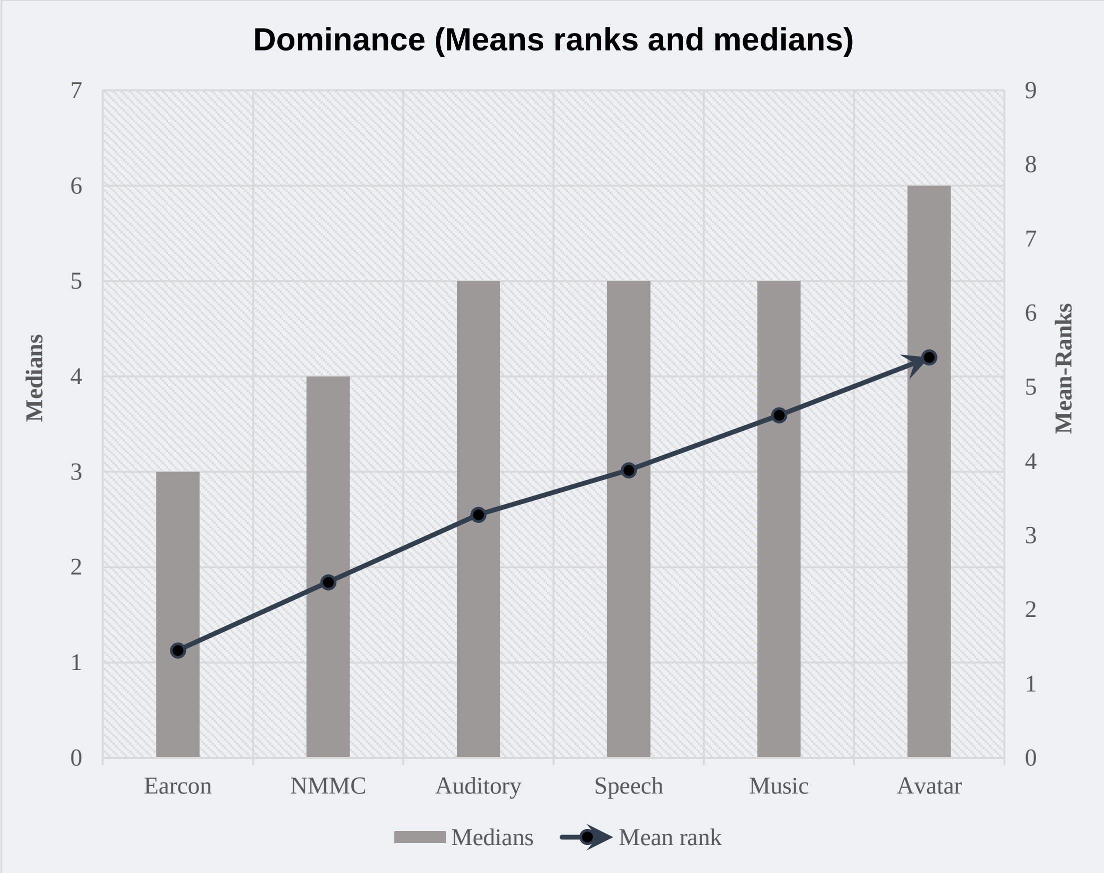


Figure 6-34: Dominance ranks.

6.5 BIOFEEDBACK

During the experimental sessions, changes in users' skin conductivity and skin temperature were monitored. The overall results are presented in Figure 6-35 and all the experimental conditions' observed skin conductivities were higher than those in the control condition were (M=6.57 µSiemens). The highest mean value of skin conductivity was obtained in the earcons condition (M=8.74 µSiemens), followed by the avatar condition (M=8.49µSiemens), the auditory icons condition (M=7.78 µSiemens), then the speech (M=7.78 µSiemens) and the music conditions (M=7.75 µSiemens). It is proven in the literature that skin conductivity linearly correlates with emotional arousal[102] [129], therefore, conditions with higher skin conductivity were more stimulating to users. With regards to skin temperature (see Figure 6-35), the highest observed mean value of skin temperature was recorded in the earcons condition (M=34.87 °C) followed by the auditory and the speech conditions (M=34.58 °C), then the avatar condition (M=34.55 °C) and lastly the music and the control condition (NNMC) (M=34.47 °C).

In general, positive emotions are not accompanied by significant increase in peripheral skin temperature. Negative emotions, especially anger are accompanied by significant increase in skin temperature; furthermore, skin temperature can be used to differentiate between negative emotions where anger for instance anger results in higher skin temperature when compared to sadness or fear. Accordingly, the differences in skin temperature are interpreted in relation to other physiological and subjective measurements. Taking into account the SAM evaluation results (see Figures 6-2, 6-5, 6-9, 6-13, 6-17 and 6-21) in addition to the skin conductivity results.

In comparison to the NMMC, the earcons condition was highly negative and the auditory icons condition was slightly negative, while the avatar, speech and music conditions were more encouraging and positively affected users' learning experience with the experimented conditions.

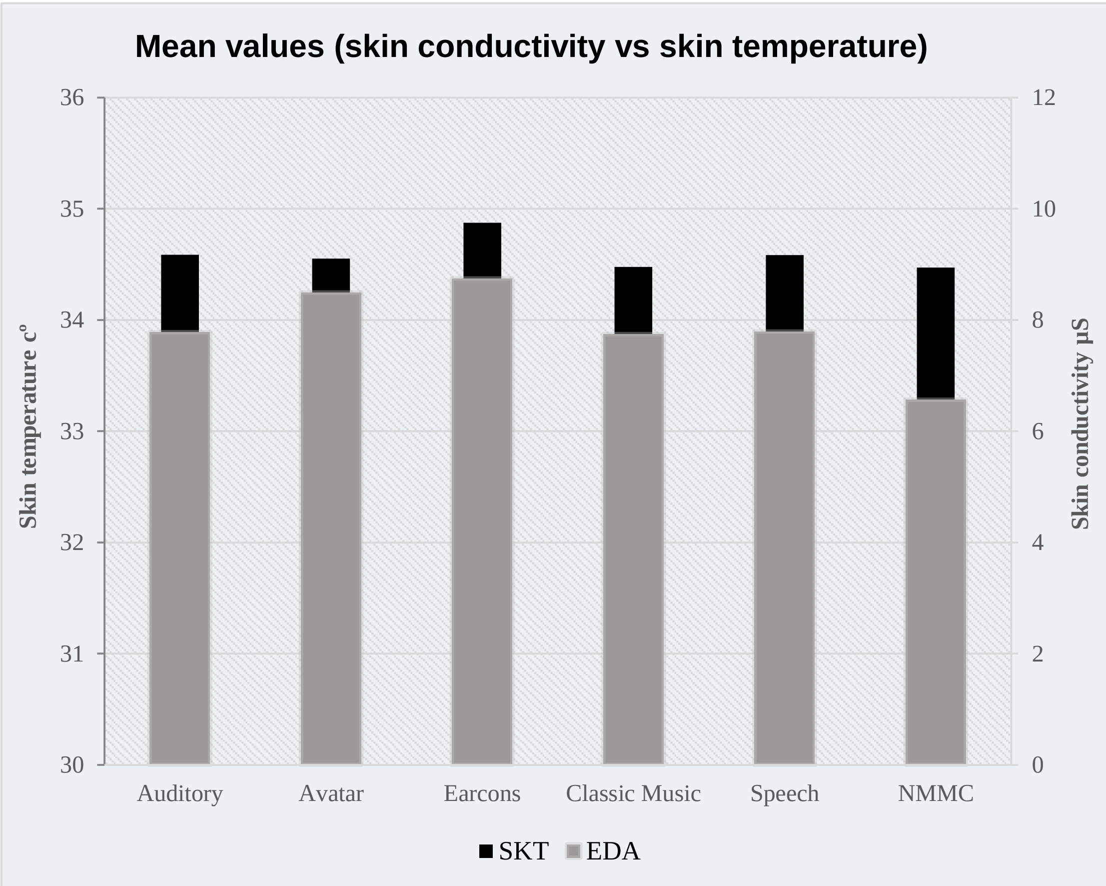


Figure 6-35: Mean values of skin conductivity and temperature.

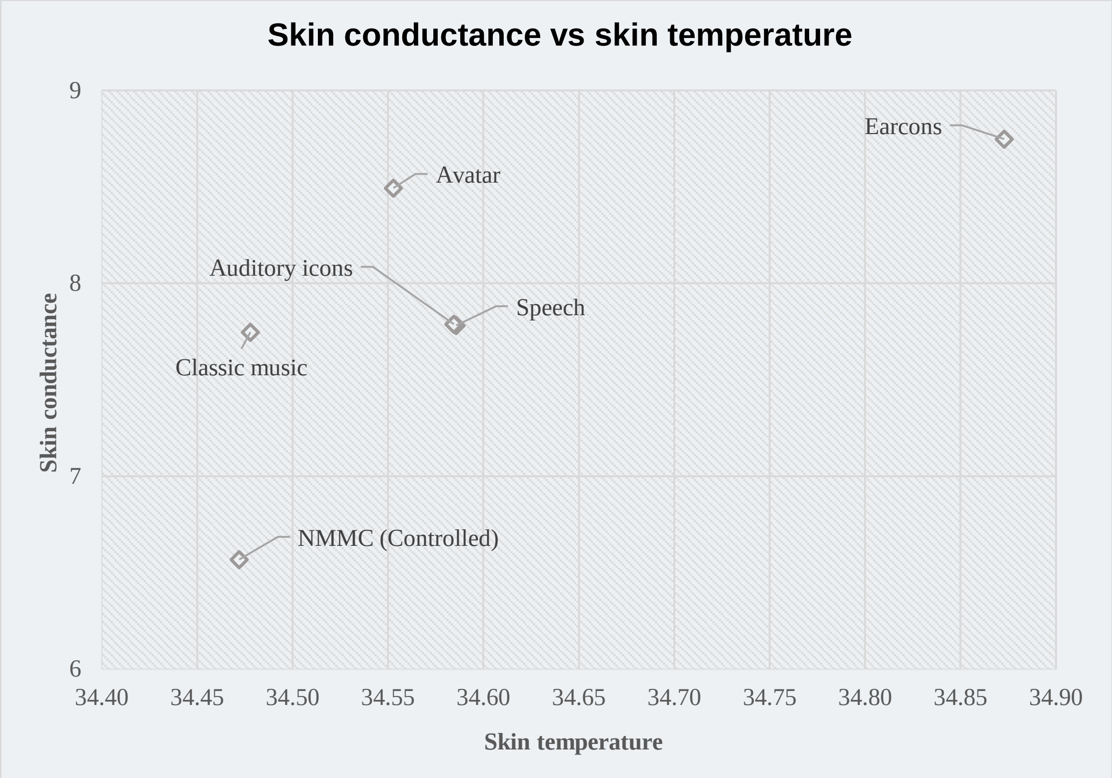


Figure 6-36: Scatterplots (skin conductivity and temperature).

6.5.1 Statistical Data Analysis

To investigate the results further and determine whether the observed differences are statistically significant, a statistical data analysis was conducted using "SPSS2013" and the results are presented below.

6.5.1.1 *Skin Conductance*

Skin conductance data was tested for normality and the results are in Table 6-4. The data is not normally distributed, therefore a non-parametric (Freidman's ANOVA) test was used to examine the differences between the experimented conditions and the results are presented in Figure 6-37. The Freidman's ANOVA test result ($\chi2(5) = 91.158$, $p < .0005$) revealed that, there is significant difference between at least two of the experimented conditions in terms of their influence on the users' arousal level, which was reflected as an increase or decrease in the users' skin conductivity. A pairwise-comparison was conducted to learn exactly where the differences lies (the results are presented in Figure 6-38 and the differences are denoted in gray). To understand the pairwise comparison results and make them clearer, the calculated means rank and medians are depicted in Figure 6-39 in an ascending order.

Tests of normality						
Condition	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory icons	.125	30	.200*	.938	30	.083
Avatar	.163	30	.041	.922	30	.031
Earcons	.169	30	.028	.925	30	.035
Music	.156	30	.060	.922	30	.031
Speech	.163	30	.040	.928	30	.042
NMMC	.146	30	.100	.929	30	.046
*. This is a lower bound of the true significance.						
a. Lilliefors Significance Correction						

Table 6-4: Normality test (skin conductance data).

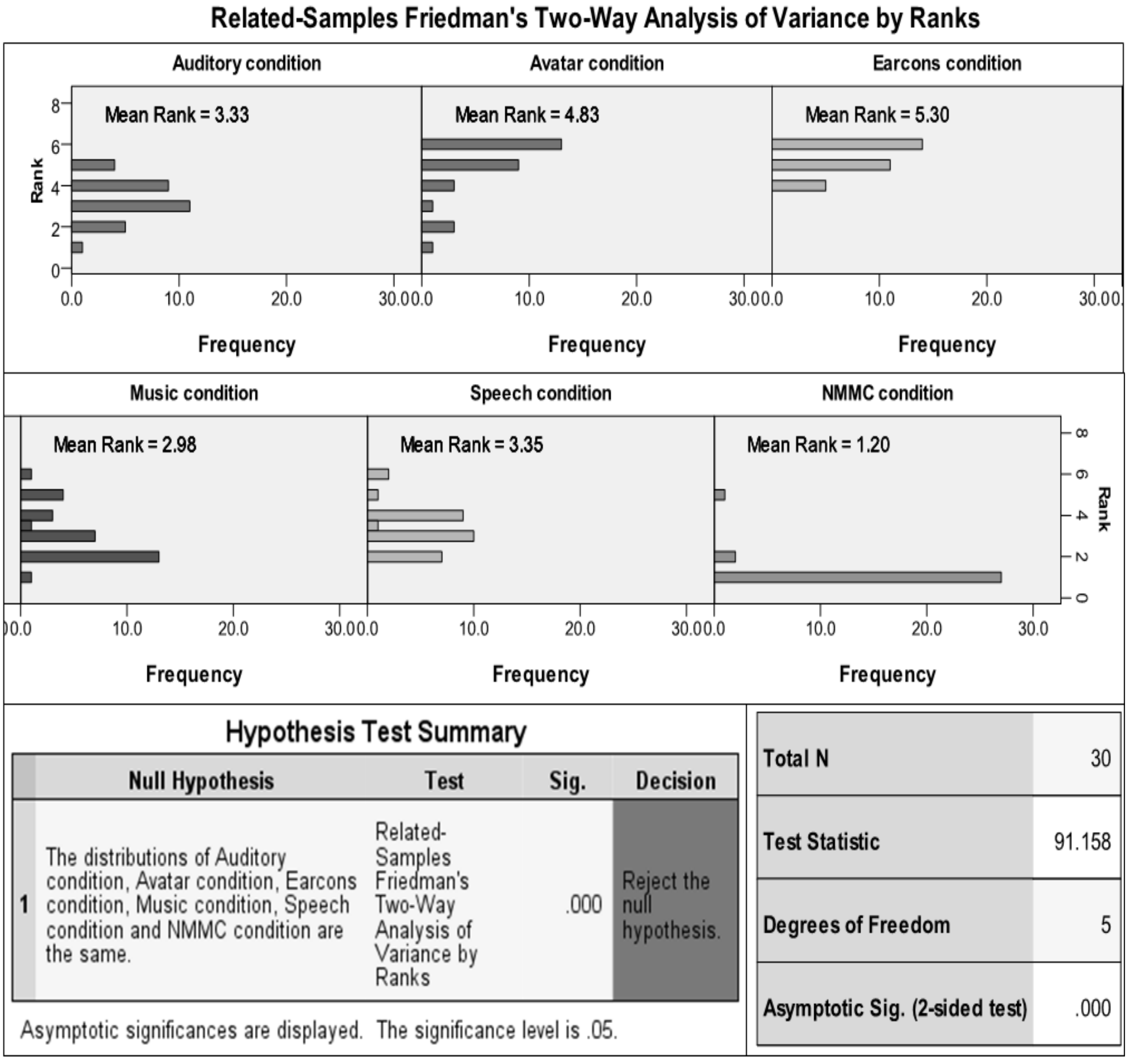
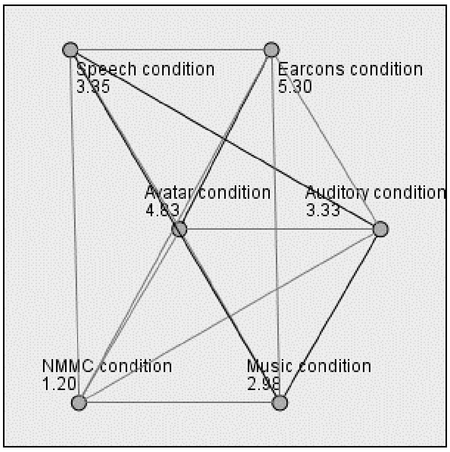


Figure 6-37: Friedman's ANOVA test (skin conductance).

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
NMMC condition-Music condition	1.783	.483	3.692	.000	.003
NMMC condition-Auditory condition	2.133	.483	4.416	.000	.000
NMMC condition-Speech condition	2.150	.483	4.451	.000	.000
NMMC condition-Avatar condition	3.633	.483	7.522	.000	.000
NMMC condition-Earcons condition	4.100	.483	8.488	.000	.000
Music condition-Auditory condition	.350	.483	.725	.469	1.000
Music condition-Speech condition	-.367	.483	-.759	.448	1.000
Music condition-Avatar condition	1.850	.483	3.830	.000	.002
Music condition-Earcons condition	2.317	.483	4.796	.000	.000
Auditory condition-Speech condition	-.017	.483	-.035	.972	1.000
Auditory condition-Avatar condition	-1.500	.483	-3.105	.002	.029
Auditory condition-Earcons condition	-1.967	.483	-4.071	.000	.001
Speech condition-Avatar condition	1.483	.483	3.071	.002	.032
Speech condition-Earcons condition	1.950	.483	4.037	.000	.001
Avatar condition-Earcons condition	-.467	.483	-.966	.334	1.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6-38: Skin conductance pairwise comparisons.

Figure 6-39 shows that the highest level of skin was measured within the avatar and the earcons with no statistical significant differences between both of them, however there are significant differences between them and the rest of the conditions. Next in rank was the auditory icons condition followed by the speech and music conditions with no statistical significant differences among them; however, the observed differences between them and the rest of the conditions are statistically significant.

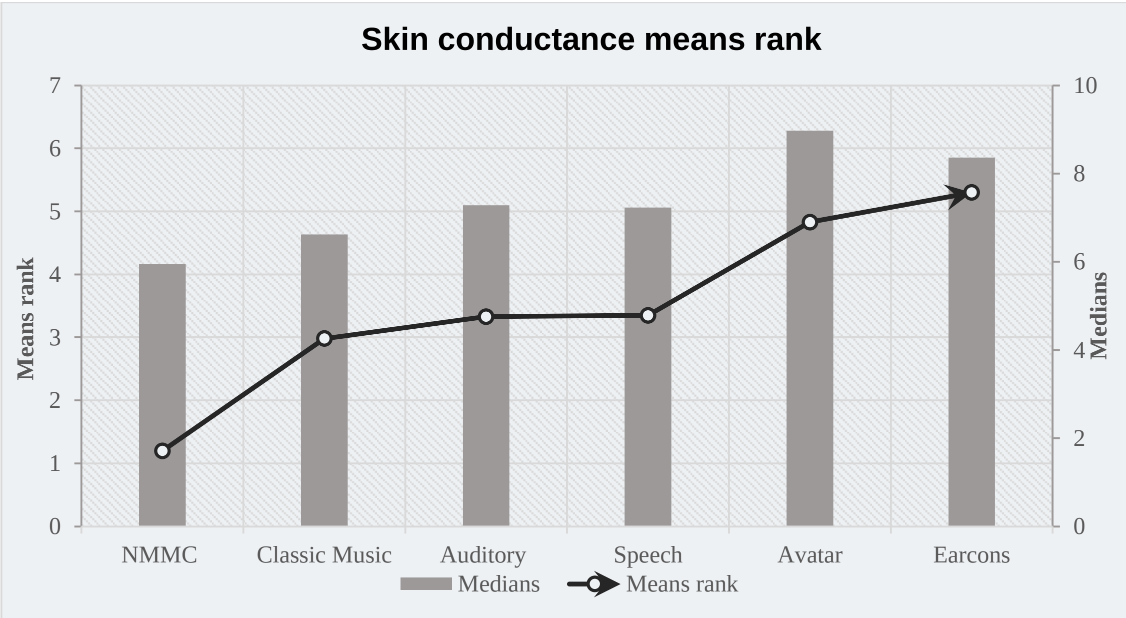


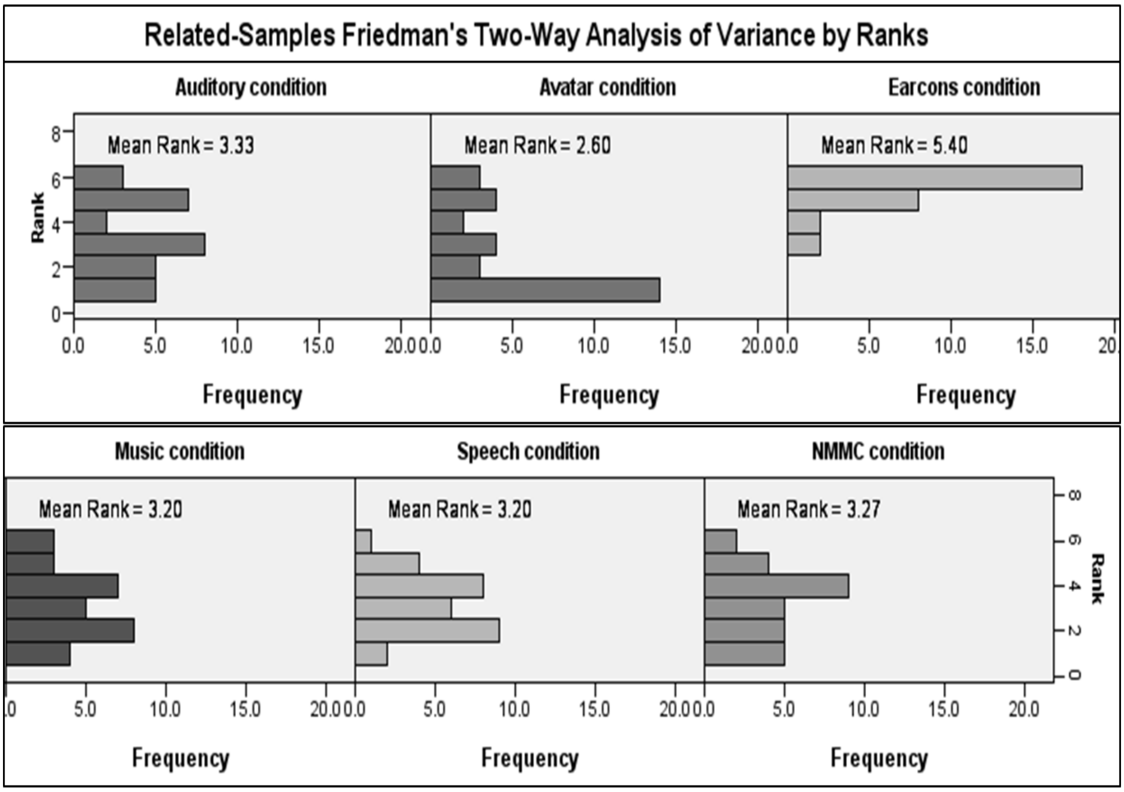
Figure 6-39: Skin conductance means rank.

6.5.1.2 *Skin Temperature*

Skin temperature data was tested for normality and the results are presented in Table 6-5. The data is not normally distributed; therefore a non-parametric (Freidman's ANOVA) test was used to examine the differences between the experimented conditions and the results are presented in Figure 6-40. The Freidman's ANOVA test result ($\chi2(5) = 40.133$, $p < .0005$.) showed that, there is significant statistical difference between at least two of the tested conditions in terms of their impact on the users' emotional valence, which is echoed as alteration in the users' measured skin temperature. A pairwise-comparison was conducted to learn where the differences existed (the results are presented in Figure 6-41 and difference are denoted in gray). To understand the pairwise comparison results and make them stronger, the calculated means rank and medians are portrayed in Figure 6-42 in ascending order.

Tests of normality						
Condition SKT	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Auditory condition	.141	30	.134	.925	30	.037
Avatar condition	.137	30	.156	.946	30	.134
Earcons condition	.170	30	.027	.918	30	.024
Music condition	.144	30	.113	.940	30	.092
Speech condition	.105	30	.200*	.940	30	.092
NMMC condition	.120	30	.200*	.950	30	.168
*. This is a lower bound of the true significance.						
a. Lilliefors Significance Correction						

Table 6-5: Skin temperature normality test.



Hypothesis Test Summary

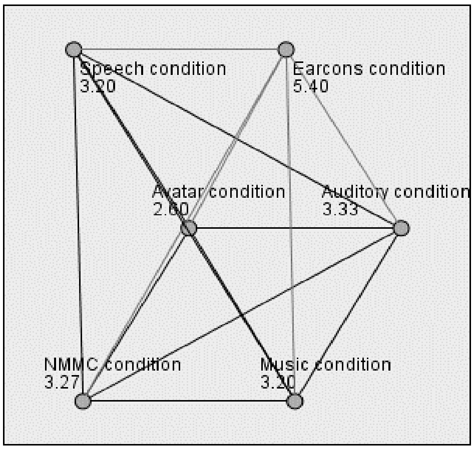
	Null Hypothesis	Test	Sig.	Decision
1	The distributions of Auditory condition, Avatar condition, Earcons condition, Music condition, Speech condition and NMMC condition are the same.	Related-Samples Friedman's Two-Way Analysis of Variance by Ranks	.000	Reject the null hypothesis.

Asymptotic significances are displayed. The significance level is .05.

Total N	30
Test Statistic	40.133
Degrees of Freedom	5
Asymptotic Sig. (2-sided test)	.000

Figure 6-40: Freidman's ANOVA test (skin tempreature).

Pairwise Comparisons



Each node shows the sample average rank.

Sample1-Sample2	Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj.Sig.
Avatar condition-Music condition	-.600	.483	-1.242	.214	1.000
Avatar condition-Speech condition	-.600	.483	-1.242	.214	1.000
Avatar condition-NMMC condition	-.667	.483	-1.380	.168	1.000
Avatar condition-Auditory condition	.733	.483	1.518	.129	1.000
Avatar condition-Earcons condition	-2.800	.483	-5.797	.000	.000
Music condition-Speech condition	.000	.483	.000	1.000	1.000
Music condition-NMMC condition	-.067	.483	-.138	.890	1.000
Music condition-Auditory condition	.133	.483	.276	.783	1.000
Music condition-Earcons condition	2.200	.483	4.554	.000	.000
Speech condition-NMMC condition	-.067	.483	-.138	.890	1.000
Speech condition-Auditory condition	.133	.483	.276	.783	1.000
Speech condition-Earcons condition	2.200	.483	4.554	.000	.000
NMMC condition-Auditory condition	.067	.483	.138	.890	1.000
NMMC condition-Earcons condition	2.133	.483	4.416	.000	.000
Auditory condition-Earcons condition	-2.067	.483	-4.278	.000	.000

Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same.
Asymptotic significances (2-sided tests) are displayed. The significance level is .05.

Figure 6-41: Pairwise comparisons (skin temperature).

To find out where the difference lies, a pairwise comparison was performed and the results are demonstrated in Figure 6-41 (the differences are designated in gray). With reference to the test result (H5) is rejected. The pairwise comparisons between the tested conditions are presented and ranked according to their means and medians in Figure 6-42. The results indicate that the differences were significant between the earcons condition as well as the control condition and the positively viewed (avatar, music, speech) conditions.

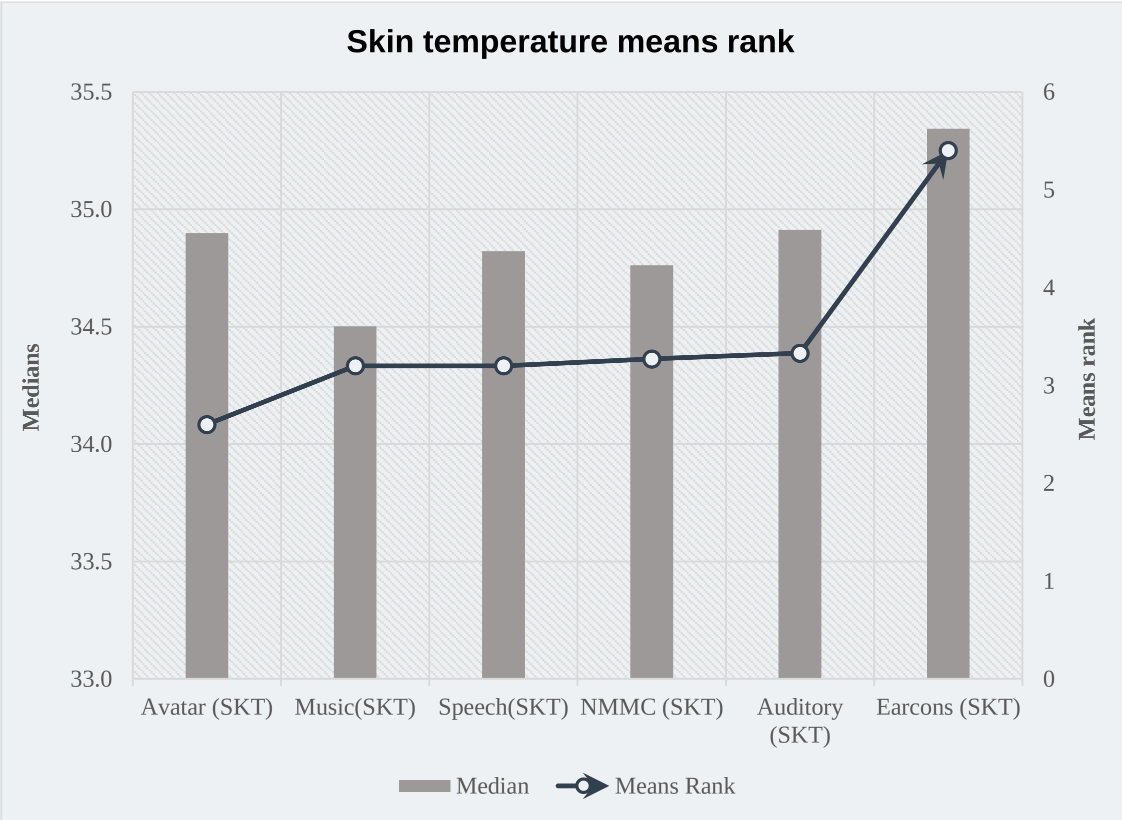


Figure 6-42: Means rank of skin temperature.

6.6 DISCUSSION

The results were obtained from users' self-assessment of their affective state in addition to the biofeedback measurements results which confirmed that the inclusion of certain communication metaphors (avatars, recorded speech and music) positively influenced the users' affective state and they emotionally moved by them. Furthermore, users felt they were neither frustrated nor dominated due to their inclusion. On the other hand, the inclusion of other interaction metaphors (earcons and auditory icons) had the opposite impact as users felt frustrated and not in control.

6.6.1 Auditory Icons Condition (induced affective state)

Figure 6-1 demonstrates user self-assessment of the condition. Users felt “slightly negative” emotion towards the condition and the strength of the emotion was neither weak nor strong. The results in show that, with the dominance evaluation, while 53.3% of the users felt “neutral”, around 40% of them felt they were to some extent not in control and they wanted to stay away from the experimented condition. Furthermore Figure 6-3 shows a triangular radar graph presenting the users’ assesment of their affective state. The triangle is not symmetrical and is deformed within the dominance and valence angles, which confirms that users’ affective state while using the auditory condition was slightly negative and they felt discouraged from continuing to use it. On the other hand, the arousal level obtained from the biofeedback measurements is shown to be stronger (M=7.78 µSiemens) in users while they were dealing with this condition. With regards to users’ skin temperature, the measured mean value (34.58 °C) for the auditory icons condition was higher than that for the control condition (34.47 °C). With reference to the subjective valence results, the increase in the measured skin tempreature can be interpreted as an indication of negative emotional experience. Overall, from the biofeedback results as well as the SAM assessment results, users encounters with auditory icons communication metaphor induced a slightly to moderately negative affective state.

6.6.2 Avatar Condition (induced affective state)

User self-assessment of the condition result is presented in Figure 6 5 and Figure 6-6. Users felt “slightly negative” emotion towards the condition and the strength of the emotion was neither weak nor strong. The results in Figure 6-7 show that, with the dominance evaluation, while 40 % of the users felt “neutral”, 60 % of them felt they were to some extent in control and they wanted to continue using the experimented condition. Figure 6-8 shows a triangular radar graph displaying the users’ assesment of their affective state. The triangle is symmetrical and the covered area within the three assessment dimensions clearly indicates that users feel pleased and positively stimulated. The result confirms that users’ affective state while using the avatar condition was highly positive and they had the desire to continue using it. Furthermore, the arousal level obtained from the biofeedback measurements show a stronger arousal level (M=8.49 µSiemens) in users while they were interacting with this condition. Concerning users’ skin temperature, the measured mean value of users skin temperature (M=34.55 °C) for the avatar condition was slightly higher than that for the control condition (34.48 °C)

Taking into account the subjective valence evaluation outcomes, the slight increase in the measured skin tempreature is seen as an indication of positive emotional expereince. In conclusion, from the biofeedback results as well as the SAM results, users' encounter with the avatar communication metaphor induced a highly positive affective state.

6.6.3 **Earcons Condition (Induced affective state)**

Figure 6-9 demonstrates users' self-assessment of the condition. They felt "slightly negative" emotion towards the condition and the strength of the emotion was neither weak nor strong. The results in Figure 6-11 show that, with the dominance evaluation, more than 97% of them felt they were to some extent dominated and they felt the urgency to stay away form the tested condition. Furthermore Figure 6-12 shows a triangular radar graph depicting the users' assesment of their affective state. The triangle is bent towards the arousal dimension, which confirms that users' affective state while using the condition was highly negative and users felt hopeless to continue using it. Furthermore, the arousal level attained from the biofeedback measurements has reveal a sturdy arousal level (M=8.74 µSiemens) in users while they encountered this condition and its communication meetaphor. Concerning the skin temperature, the measured mean value of users' skin temperature (M=34.87 °C) was higher than the measured mean value within the control condition (34.48 °C). Considering the subjective valence results, the escalation of skin tempreature is understood as an indication of negative emotional experience. Largely, from the biofeedback results as well as the SAM valuation results, users utilizing the auditory icons communication metaphor, resulted in a moderately to highly negative affective state.

6.6.4 **Classic Music Condition (Induced affective state)**

Figure 6-13 and Figure 6-14 demonstrate users' self-assessment of the condition,they felt positive emotion towards the condition and the strength of the emotion was slightly strong. The results shown in Figure 6-15 indicate the dominance evaluation. While 83.40% of the users feelings in the condition ranged from neutral to in control and had the intention to continue using it, only 13.6% of them felt they were to some extent not in control and wished to break away. Figure 6-16 shows a three-sided radar graph presenting the users' assesments off their affective state. The triangle is symmetrical and well balanced within the dominance as well as the arousal and valence angles. This indicates that users' affective state while using the classical music was balanced and moderately positive. From the biofeedback point of view, the arousal level obtained from measuring the changes in users' skin conductivity shows a slight increase (M=7.75 µSiemens). With regards to the changes in users' skin temperature, the measured mean value for the condition (34.47 °C) was almost equal to that obtained from the control condition (34.48 °C) With reference to the subjective valence results, the skin tempreature can be interpreted as an indication of the music's positive and calming effect. Overall, from both the subjective as well as the objective assessment results, users' encounters with the classical music communication metaphor resulted in a positive and calming affective state.

6.6.5 **Recorded Speech Condition (Induced affective state)**

Figures 6-17 and 6-18 display user self-assessment of the condition. They felt "neutral" to "slightly positive" emotion towards the condition and the strength of the emotion was moderate. The results in Figure 6-19 show that, within the dominance evaluation, while 76.6% of the users felt "neutral" and did not have any desire to get away from using it, in only 23.40% felt they were controlled to some extent and wanted to stay away from it. Figure 6-20 shows a three-cornered radar graph presenting the users' views of their affective state, The triangle is balanced with slight distortion within the dominance angle. The arousal level obtained from the biofeedback measurements shows moderate arousal level (M=7.79 µSiemens) in users users skin conductivity. With regards to the users' skin temperature, the measured mean value of users' skin temperature (34.58 °C) was higher than the value obtained for the control condition (34.48 °C) With reference to the subjective valence results the increase in the measured skin tempreature can be interpreted as an indication of neutral to positive emotional experience. Overall, from the biofeedback results as well as the SAM assessment results, users encounters with the classical music communication metaphor experienced a moderately positive affective state.

6.6.6 **All Conditions**

Valence level: In relation to the results obtained from the evaluation of valence dimension, with the exception of the earcons and the auditory icons communication metaphors, users' feelings towards the employed communication metaphors ranged from highly positive (avatar and music metaphors) to moderately positive (speech metaphor).

Consequently, based on the statistical tests in Section 6.4.8.1 (see Figure 6-26) and its proceeding post-hoc test (see Figure 6-27 and Figure 6-28), H1 is rejected.

Arousal level: Within the arousal dimension in SAM results, the most stimulating communication metaphor was the avatar communication metaphor, followed by the earcons then the classic music metaphors. The results confirmed that, all three are affectively strong and their ability to stimulate users is high. Next, in the rank within the arousal dimension are the speech and auditory icons; users were to some extent stimulated by them and they outperformed the control condition in this users' affective state dimension. The statistical test results in Section 6.4.8.2 (see Figure 6-29) and its proceeding *post-hoc* test (see Figure 6-30 and Figure 6-31) indicated that the utilized communication metaphors were not equally affecting. Consequently based on the aforementioned results, what was hypothesised in H2 is rejected.

Dominance level: Within the SAM dominance measurements, users' ratings ranged from feeling in-control (avatar, music and speech communication metaphors) to feeling slightly controlled (auditory icons metaphor) to strongly controlled (earcons metaphor). The results in the dimension were statistically tested (refer to Section 6.4.8.3 and Figure 6-32) and compared for all of the tested conditions (Figure 6-33 and Figure 6-34); they confirmed that the differences between them were statistically significant. The *post-hoc* tests using pairwise comparisons proved that certain interaction metaphors were superior to others. Consequently, what was hypothesized in H3 is rejected.

Biofeedback: Concerning the biofeedback evaluation outcomes, the results presented in Figure 6-35, show that users' skin conductivity was affected by the interaction metaphors employed. The observed changes varied based on the type of communication metaphor in use; an increase in highest skin conductivity was observed with the earcons communication, followed by the avatar and auditory conditions communication metaphor condition, while the remaining communication metaphors were comparably equal. Skin conductivity is considered as a strong marker for detecting the levels of emotional arousal. The results show that it was more accurate than SAM subjective assessment in this regard. It seems users' evaluation of their arousal was actually overstated due to the dominance of the experimented communication metaphors and their feeling of being either controlled or in control. Based on the performed statistical tests (see Section 6.5.1.1, Figures 6-38 and 6-39) what was hypothesised in H4 is rejected. With reference to skin temperature measurements, the highest increase in skin temperature was observed in the earcons condition, followed by the auditory icons and speech conditions. In the next level was the avatar condition followed by the classical music condition. Normally, changes in peripheral skin temperature in relation to users; affective state are difficult to interpret, as such increases might occur due to either negative or positive valence. Nonetheless, in the case of negative valence it can actually be used to differentiate sadness from anger, which is accompanied by higher peripheral skin temperature; thus, they are inferred with the subjective measurements and interpreted accordingly. Accordingly the increase seen accompanying both earcons and auditory icons communication metaphors is an indicator of negative emotions, which is anger in the case of the earcons communication metaphor and sadness with regards to the auditory icons. The results in this category are comparable with those obtained in the SAM assessments. Based on the results obtained as well as executed statistical tests (see Section 6.5.1.2, Figures 6-40, 6-41 and 6-42) what was hypothesised in H5 is rejected. In recognition of the results obtained from biofeedback and SAM measurements and the executed statistical tests, what was proposed in hypotheses H6 is rejected. Overall, with reference to the results obtained from the SAM self-assessment test (see comparisons in Figure 6-25) and users' biofeedback measurements (see comparisons in Figure 6-36) it is possible to conclude that, communication metaphors had a profound influence on the users' affective state. Accordingly, they can be divided into two categories: positive (avatar, classic music and recorded speech) and negative (earcons and auditory icons). Furthermore, the results clearly show that using earcons as well as auditory icons can be risky due to their acoustic strength and the influence that might have on the users' affective state.

6.7 SUMMARY

Users' affective state was evaluated while interacting with experimental conditions using objective and subjective measurements. The results confirm that the inclusion of interaction metaphors could have a strong constructive or destructive impact on learners' affective states. The inclusion of speaking affective avatar as well as classical music and recorded speech communication metaphors had a positive impact on the users' affective state and they should be employed in e-learning systems. While both the earcons and the auditory icons communication metaphor had a negative impact upon users' affective state, the results show that they are a strong acoustic tool, which can be utilized cautiously with consideration in alerting and making the learning environment more realistic. Nevertheless, when communication metaphors were utilized in multimodal metaphoric communication setup in the first experimental phase (3.8), they performed better and their side effects were minimized. The overall assessment results are presented in the affective state traffic light below (see Figure 6-43).

	User afective satae trafic light		
Condition	**SAM**	**Biofeedback**	**Affective state**
Auditory	✗	!	!
Avatar	✓	✓	✓
Earcons	✗	✗	✗
Music	✓	✓	✓
Speech	✓	✓	✓
NMMC	!	!	!

Figure 6-43: Affective state traffic light.

CHAPTER 7

7 EMPIRICALLY DERIVED GUIDELINES AND RECOMMENDATIONS

The aim of this research was to evaluate the overall e-learning user experience with multimodal communication and multimodal e-learning interfaces. Furthermore, it aimed to investigate the impact of particular multimodal communication metaphors on multimodal e-learning interfaces, using a triple evaluation approach consisting of usability measurements (effectiveness, efficiency, user satisfaction and learning performance), in addition to user experience measurements (hedonic quality, pragmatic quality and attractiveness) and Affective state measurements (biofeedback and SAM).

In order to fulfil these aims and to answer the research questions in Section 1.2, two experimental phases were designed and executed. The first experimental phase showed that using the multimodal condition, which included the following communication metaphors: "auditory icons, earcons, facially expressive avatars, classical music, recorded speech, text and graphics" scored better than the non-multimodal condition in terms of system usability, user experience and affective state (refer to Chapter 3 for details). A further investigation was needed to highlight the impact of the individual interaction metaphors and their contribution to the overall observed improvements in the first experimental stage. To achieve the anticipated goal, a second experimental phase was executed with six conditions, five of which were dominated by only one interaction metaphor, while the sixth condition was a non-multimodal (controlled), similar to that in the first experimental phase. The results showed that the avatar made the highest contribution to the achieved performance results followed by speech and classical music, and then auditory icons and earcons (refer to Chapter 4 for usability evaluation, Chapter 5 for user experience evaluation and Chapter 6 for user affective state evaluation).

7.1 CONCLUSION

Based on the results obtained in this experimental work and the adopted triple evaluation approach, a series of empirically derived guidelines and recommendation were made.

1. Use of auditory icons, based on the results of the empirical work and the results obtained from the evaluation of the impact of using auditory icons on usability (Section 4.5.10),

user experience (Section 5.4), biofeedback (Section 6.5) and SAM (Section 6.4) of the tested condition. The inclusion of auditory icons helped to make the learning environment more realistic and stimulated users. Users' experience with the metaphor was slightly positive, especially within the hedonic quality and attraction dimensions. However, there was no noticeable difference within the aspects of pragmatic quality and usability. The results showed that users were, to some extent, satisfied with the use of auditory icons, however the interface efficiency as well as effectiveness and learning performance were not improved by its inclusion. Consequently, it is recommended that auditory icons should be incorporated into multimodal e-learning interface with care when their use is essential. Furthermore, with reference to the results of the first experimental phase, auditory icons should be used in conjunction with other interaction metaphors. The extensive use of auditory icons is not recommended as it might hinder the interface usability, induce a negative affective state and result in an unpleasant user experience.

2. Use of avatars, based on the results of the empirical work and the results obtained from the evaluation of the impact of using facially expressive avatars on usability (Section 4.5.10) in addition to user experience (Section 5.4) as well as SAM (Section 6.4) and biofeedback (Section 6.5) of the tested condition. The results show that this interaction metaphor provided highly pragmatic as well as hedonic and attractive user experience. Within the usability assessment criteria, users were highly satisfied with it; furthermore, it improved the interface efficiency and effectiveness and increased users' learning performance. Furthermore, users were less concerned about missing human-to-human communications with the presence of avatars. Concerning users' affective states, they were stimulated by the inclusion of this interaction metaphor, which caused a positive valence and made users feel pleased and motivated. For these reasons, it is highly recommended that facially expressive avatars be used whenever possible, due to their positive impact on overall e-learning user experience.
3. Use of Earcons, based on the results of the empirical work and the results obtained from the evaluation of the impact of using earcons on usability (Section 4.5.10), user experience (Section 5.4) and biofeedback (Section 6.5) and SAM (Section 6.4.3) of the tested condition, the inclusion of earcons had a negative impact on the learning environment. Despite the fact that users were alerted by the metaphor, they were annoyed, stressed and demotivated. Users' experience was strongly affected within its pragmatic and hedonic qualities and they were not attracted to it. Furthermore, with

reference to the usability elements (efficiency, effectiveness satisfaction, and learning performance) the presence of earcons was severely negative and deteriorated users' learning performance. In terms of users' affective state, the earcons metaphor had a negative valence with strong arousal, which was reflected in SAM as well as biofeedback results. Accordingly, it is recommended that earcons should be integrated into multimodal e-learning interfaces with caution as they have a strong influence on users' affective state. They should be used only when it is necessary to alert or direct the user; moreover, with reference to the results of the first experimental phase, earcons should be used in combination with other interaction metaphors. Overusing earcons in interfaces could have severe side-effects and might deter the overall user experience and performance.

4. Use of music, based on the results of the empirical work and the results obtained from the evaluation of the impact of using classical music on usability (see Section 4.5.10), user experience (Section 5.4) and biofeedback (Section 6.5) and SAM (Section 6.4.4) of the tested condition. The results indicate that the inclusion of classical music interaction metaphor helped the users to relax and engage with the presented topic. The interaction metaphor noticeably enhanced usability in terms of effectiveness, efficiency and user satisfaction. Furthermore, the learning performance was improved and the absence of human-to-human communication was less noticeable. Concerning the user experience dimensions, the metaphor improved the hedonic and pragmatic qualities as well as attractiveness. Concerning the users' affective states, the inclusion of the metaphor resulted in a positive valence, and stimulated and motivated users. Consequently, it is recommended that this metaphor be used due to its positive impact on the overall users' experience and learning environment.
5. Use of recorded speech, based on the results of the empirical work and the results obtained from the evaluation of the impact of using recorded speech on usability (see Section 4.5.10), user experience (Section 5.4) as well as biofeedback (Section 6.5) and SAM results (Section 6.4) of the tested condition. The obtained results indicated that the inclusion of the speech interaction metaphor enhanced the learning environment and was positively viewed. In terms of user experience, the metaphor improved the pragmatic as well as the hedonic qualities and was attractive to users. Within the affective state assessments, the metaphor resulted in a positive valence and slightly stimulated the users and the overall impression ranged between neutral and positive. Within usability evaluation, the observed impact of the metaphor on

efficiency, effectiveness and users' satisfaction was positive. Users' engagement and learning performance were improved. With reference to the above-mentioned results, the use of this interaction metaphor is highly recommended due to its ability to enhance the e-learning environment as well as the overall user experience, especially when demonstrating or requiring user attention.

6. Communication metaphors and frequency of use, with reference to the results obtained from this, empirical work within the usability assessment dimensions (Section 4.6), the user experience dimension (Section 5.4) and the users' affective state dimension (Section 6.4 and 6.5). The results show that auditory icons and earcons communication metaphors can be hazardous especially when used on their own, their usage should be minimal and it is preferred that they are combined with the other communication metaphors such as recorded speech and avatar. In accordance with the results obtained in the first experimental phases (Section 3.6), that approach will minimize their side effects and bring more benefit in terms of system usability, provided user experience as well as their learning performance and positively influencing the users' affective states.

7.2 REFLECTION

The results obtained from this empirical work clearly shows that, adopting a more sophisticated evaluation approach that goes beyond the conventional concept of usability testing can uncover the shortcomings and pitfalls of the system. This is crucial to the success of the system, especially in the e-learning field where users' experience and system provided pragmatic as well as hedonic qualities as well as attractiveness are decisive for long-term adaption. The proposed triple evaluation approach that included usability evaluation, user experience evaluation and affective state evaluation could help system developers to successfully meet their targets and address users' needs. Furthermore, the study shows that user affective state evaluation could help developers identify how users' emotions are triggered and influenced by the system and its presentation mode. Furthermore, in addition to users' self-assessment techniques, using biofeedback in particular can provide unbiased insight into users' affective state while interacting with the system and how they are affected in response to the presented materials. Concerning user experience and hedonic quality; the results obtained from the empirical assessment prove that hedonic quality is an added value to any interactive system, due to its constructive impact on how users perceived the system presented to them. The experimental conditions, which were rated by users to be high in hedonic quality, scored high in their usability and users' affective state assessments indicate that using certain communication metaphors such as avatar, classical music as well as recorded speech could improve the system image with their hedonic values. In terms of system perceived attractiveness, the results obtained from the empirical assessment prove that attractiveness is an added value to any system, due to its appositive impact on how users perceived the system presented to them. The experimental conditions, which were rated by users to be attractive, scored high in their usability and users' affective state assessments indicate that using certain communication metaphors such as avatar, classic music as well as recorded speech could improve the system image with their attractiveness. In terms of engagement and alertness, the results obtained from the empirical work show that using certain communication metaphors improved users engagement with the systems and the learning topics presented to them. The enhancements in users' engagements had a positive impact on the system effectiveness and were reflected in the user learning performance. Therefore employing communication metaphors such as avatar, classical music and recorded speech is highly recommend for improving users' engagement. Furthermore, the results show that the earcons communication metaphor can be a powerful tool for alerting users or highlighting important information. It is highly recommended that it be used it in this way, without over-using it, to minimize its side effects. In terms of convenience, the results obtained from the experimental work reveal that employing certain interaction metaphors had a positive impact on usability by making the course of interaction between the users and the system a convenient and easy experience. Accordingly, we recommend using such communication metaphors to make the users comfortable dealing with any system presented to them, as doing so will positively influence their affective state as well as their experience and attract them to using them system. Furthermore , the results obtained in both the first and the second experimental phases show that using certain communication metaphors (avatars, classic music and speech), either in singular or combined mode, could to some extent substitute human to human communication and feeling isolated. The use of objective assessment methods, such as Affectiva QSensor, which is a non-intrusive biofeedback device, was used to monitor changes in the users' skin conductance and temperature while interacting with the experimental conditions. This approach provided us with unbiased information about the users' affective state within the dimensions of arousal as well as valence and how employing certain communication metaphor influenced them; subsequently it is highly recommended that, in addition to the subjective assessment techniques, biofeedback should be included in any future system with human centred design. Based on the empirical results obtained, the triple evaluation was used for its benefit in exploring the experimented conditions in depth and revealing their pros and cons, which cannot be achieved using a single evaluation approach such as evaluating either usability or user experience and users' affective state on its own. Accordingly, we propose including the triple evaluation approach within the SDLC of any multimodal e-learning system prior to their release.

The scope of work in this research was limited to only five interaction metaphors, the setup in which they were used in this investigation. The research did not investigate other methods of delivering e-learning materials such as storytelling, edutainment, and game based learning. Furthermore, the results obtained from the biofeedback were limited because only two physiological signals were recorded. Another limitation was that the results obtained were observed and collected from first time users and did not address how long-term use could influence the perceived system usability as well as users' experience and affective states.

7.3 FUTURE WORK

In order to extend this empirical work ad overcome the areas, which were not covered in this investigation, we propose the following:

1. The use of more physiological measurements to better understand what is going on inside the users, their reactions, their emotions and how they feel about the e-learning interfaces presented to them and the interaction metaphors used to communicate the learning materials to them.
2. Utilizing the triple evaluation approach to explore other delivery methods in e- learning such as game based learning and storytelling.
3. The use of the triple evaluation approach to assess the impact of using user-customized interaction metaphors, such as avatars and earcons, on the overall e-learning user experience.

4. The use of the proposed evaluation approach in assessing haptic feedback in relation to the cognitive theory of multimedia learning.
5. Using the employed triple evaluation approach within this study to assess the impact of long-term use of multimodal interaction on the user experience and affective state.

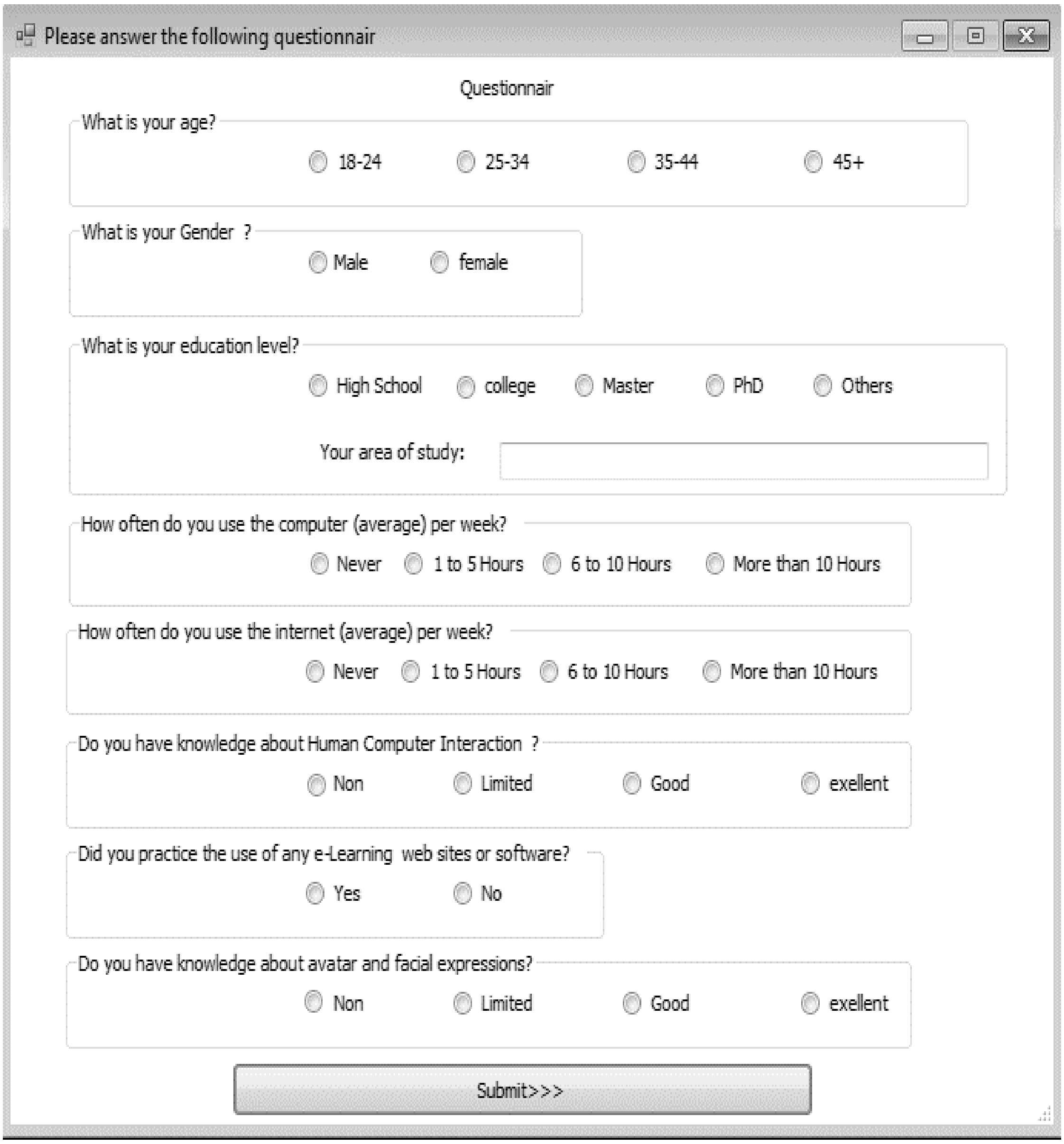
7.4 EPILOGUE

This research used a triple evaluation approach to investigate the impact of using communication metaphors such as auditory icons, avatars, earcons, classical music and recorded speech in e-learning interfaces on the overall e-learning user experience. The results show that the inclusion of such metaphors was beneficial; they were hedonic, pragmatic, attractive, efficient, effective, satisfying, and induced a positive affective state. Therefore, we can conclude that, the use of multimodal communication in e-learning interfaces improved the overall e-learning user experience. Individual interaction metaphors were not equal in their impact on users and the overall e-learning user experience. Within the three evaluation areas (usability, user experience and affective state) facially expressive speaking avatars gained the highest rank, followed by recorded speech and classical music, then auditory icons and lastly earcons. The evaluation criteria included assessing each communication metaphor in terms of their efficiency, effectiveness, user satisfaction, learning performance, hedonic quality, pragmatic quality, attractiveness, arousal, valence, dominance, changes in skin conductance and temperature. The study results indicate that there are three interaction metaphors (facially expressive speaking avatar, recorded speech and classical music) with a high potential for improving e-learning user experience, which should be extensively used in e-learning interfaces. On the other hand, regarding the remaining metaphors, auditory icons should be used only when it is essential, while earcons are very affective, therefore their use should be minimized and utilized with great care. Furthermore, using earcons as well as auditory icons in multimodal scenario in conjunction with other communication metaphors could minimize or eliminate their negative side effects. The results obtained from the proposed evaluation approach, strongly suggest that user experience and affective state evaluations should be included in addition to usability evaluation within SDLC of multimodal e-Learning systems.

APPENDICES

APPENDIX A
PRE AND POST EXPERIMENTS SHARED ITEMS

Pre-session questionnaires


Please answer the following questionnair

Questionnair

What is your age?
18-24 | 25-34 | 35-44 | 45+

What is your Gender ?
Male | female

What is your education level?
High School | college | Master | PhD | Others

Your area of study:

How often do you use the computer (average) per week?
Never | 1 to 5 Hours | 6 to 10 Hours | More than 10 Hours

How often do you use the internet (average) per week?
Never | 1 to 5 Hours | 6 to 10 Hours | More than 10 Hours

Do you have knowledge about Human Computer Interaction ?
Non | Limited | Good | exellent

Did you practice the use of any e-Learning web sites or software?
Yes | No

Do you have knowledge about avatar and facial expressions?
Non | Limited | Good | exellent

Submit>>>

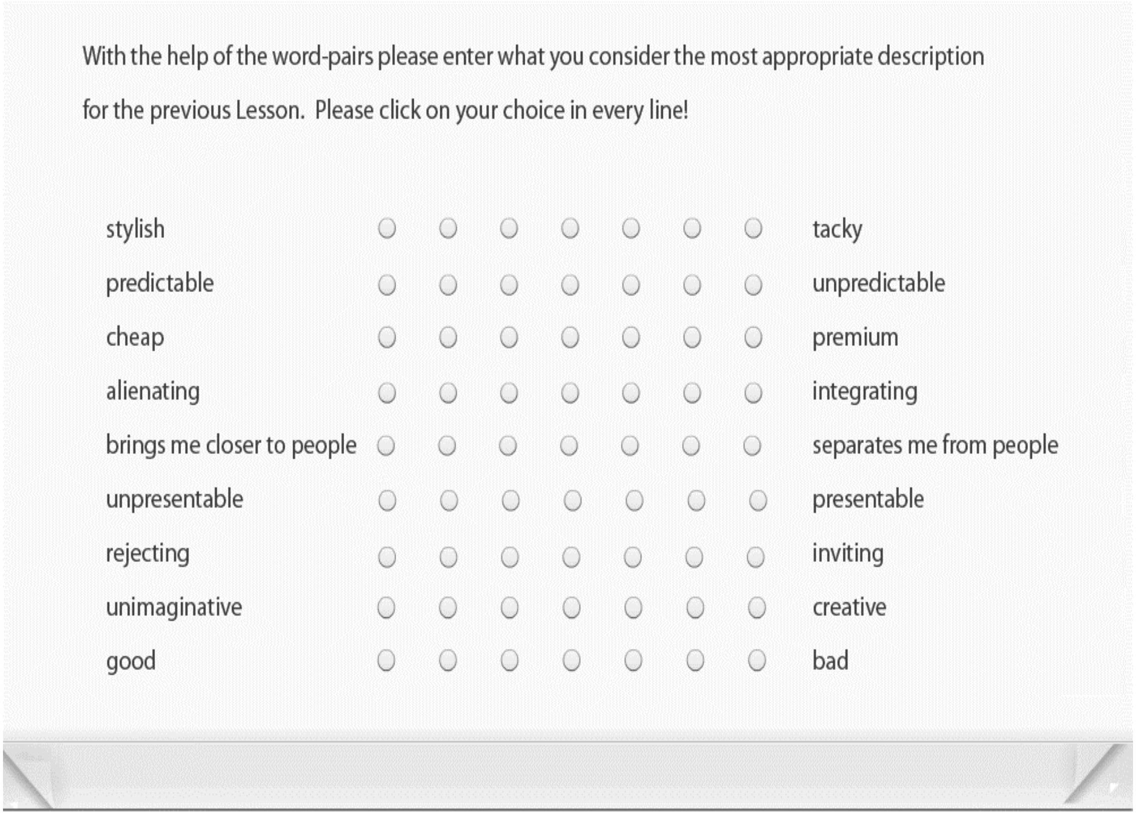
AttrakDiff (user experience evaluation instrument)

With the help of the word-pairs please enter what you consider the most appropriate description for the previous Lesson. Please click on your choice in every line!

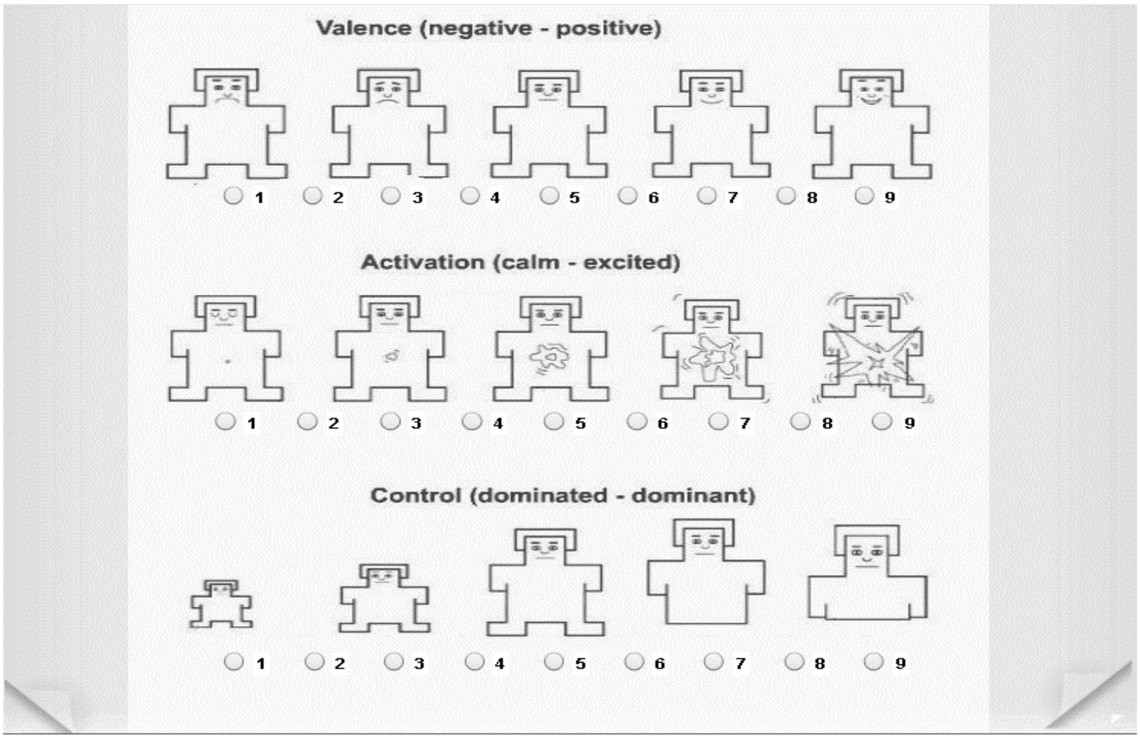
human	○	○	○	○	○	○	○	technical
isolating	○	○	○	○	○	○	○	connective
pleasant	○	○	○	○	○	○	○	unpleasant
inventive	○	○	○	○	○	○	○	conventional
simple	○	○	○	○	○	○	○	complicated
professional	○	○	○	○	○	○	○	unprofessional
ugly	○	○	○	○	○	○	○	attractive
practical	○	○	○	○	○	○	○	impractical
likeable	○	○	○	○	○	○	○	disagreeable
cumbersome	○	○	○	○	○	○	○	straightforward

With the help of the word-pairs please enter what you consider the most appropriate description for the previous Lesson. Please click on your choice in every line!

confusing	○	○	○	○	○	○	○	clearly structured
repelling	○	○	○	○	○	○	○	appealing
bold	○	○	○	○	○	○	○	cautious
innovative	○	○	○	○	○	○	○	conservative
dull	○	○	○	○	○	○	○	captivating
undemanding	○	○	○	○	○	○	○	challenging
motivating	○	○	○	○	○	○	○	discouraging
novel	○	○	○	○	○	○	○	ordinary
unruly	○	○	○	○	○	○	○	manageable



SAM user interface



Biofeedback (SC and SKT) log file sample

```
Log File Created by Q V1.0 - (c) 2010 Affectiva Inc.
File Version: 1.01
Firmware Version: 1.50
UUID: AFQ321100YW
Start Time: 2012-03-09 14:10:22 Offset:-00
 Z-axis | Y-axis | X-axis | Battery | °Celsius | EDA(uS)
---------------------------------------------------------
-0.360,-0.500,1.060,-1,26.000,1.905
-0.370,-0.630,1.090,-1,26.000,1.901
-0.410,-0.890,0.850,-1,26.000,1.889
-0.140,-0.540,1.100,-1,26.000,1.885
-0.330,-0.810,0.710,-1,26.000,1.889
-0.390,-0.740,0.990,-1,26.000,1.889
-0.350,-0.800,0.840,-1,26.000,1.901
-0.350,-0.840,0.700,-1,26.000,1.885
-0.350,-0.880,0.630,-1,26.000,1.870
-0.330,-0.960,0.720,-1,26.000,1.859
-0.290,-0.980,0.570,-1,26.000,1.855
-0.270,-0.960,0.320,-1,26.000,1.855
-0.290,-1.020,0.470,-1,26.000,1.852
-0.310,-1.070,0.500,-1,26.000,1.870
-0.260,-0.990,0.430,-1,26.000,1.859
-0.280,-1.100,0.430,-1,26.000,1.844
-0.290,-1.000,0.280,-1,26.000,1.829
-0.250,-1.080,0.350,-1,26.000,1.825
-0.260,-1.020,0.320,-1,26.000,1.810
-0.210,-0.980,0.280,-1,26.000,1.796
-0.340,-1.060,0.260,-1,26.000,1.781
-0.300,-1.140,0.320,-1,26.000,1.781
-0.280,-1.000,0.200,-1,26.000,1.781
-0.290,-1.060,0.260,-1,26.000,1.767
-0.260,-1.130,0.210,-1,26.000,1.767
-0.290,-1.020,0.160,-1,26.000,1.752
-0.360,-0.970,0.270,-1,26.000,1.756
-0.310,-1.070,0.360,-1,26.000,1.767
-0.290,-1.010,0.430,-1,26.000,1.767
```

APPENDIX B

FIRST EXPERIMENT

First experiment: pre-session questionnaire data

User	Age	Gender	Education	Area Of Study	Use of computer/week:	Use of Internet/week:	Knowledge about earthquake:	Knowledge about Avatar and facial expressions:	Practice or use of eLearning websites and software:
1	18-24	Female	Undergraduate	Health	> 10 Hours	> 10 Hours	Non	Limited	Yes
2	45-54	Male	PhD	Engineering	6-10 Hours	6-10 Hours	Excellent	Non	Yes
3	25-34	Male	PhD	Computing	> 10 Hours	> 10 Hours	Non	Excellent	Yes
4	35-44	Male	PhD	Computing	> 10 Hours	> 10 Hours	Non	Excellent	Yes
5	35-44	male	PhD	Engineering	> 10 Hours	> 10 Hours	Excellent	Limited	Yes
6	25-34	female	MSc	Art	> 10 Hours	> 10 Hours	Non	Good	Yes
7	25-34	female	MSc	Language	> 10 Hours	> 10 Hours	Non	Non	Yes
8	25-34	male	Msc	Law	> 10 Hours	> 10 Hours	Limited	Non	Yes
9	18-24	female	Undergraduate	Health	> 10 Hours	> 10 Hours	Non	Non	Yes
10	18-24	male	Undergraduate	Media	> 10 Hours	> 10 Hours	Limited	Good	Yes
11	25-34	male	MSc	Law	> 10 Hours	> 10 Hours	Non	Limited	Yes
12	25-34	male	Undergraduate	Engineering	> 10 Hours	> 10 Hours	Limited	Non	Yes
13	25-34	male	MSc	Computing	> 10 Hours	> 10 Hours	Limited	Limited	Yes
14	25-34	female	MSc	Health	> 10 Hours	> 10 Hours	Non	Non	Yes
15	18-24	male	Undergraduate	Art	> 10 Hours	> 10 Hours	Non	Good	Yes
16	18-24	male	Undergraduate	Engineering	> 10 Hours	> 10 Hours	Good	Non	Yes
17	25-34	male	MSc	Media	> 10 Hours	> 10 Hours	Non	Good	Yes
18	18-24	female	Undergraduate	Art	> 10 Hours	> 10 Hours	Non	Limited	Yes
19	18-24	female	Undergraduate	Business	> 10 Hours	> 10 Hours	Non	Limited	Yes
20	18-24	female	Undergraduate	Business	> 10 Hours	> 10 Hours	Limited	Limited	Yes
21	18-24	male	Undergraduate	Health	> 10 Hours	> 10 Hours	Non	Non	Yes
22	18-24	male	Undergraduate	Computing	> 10 Hours	> 10 Hours	Non	Limited	Yes
23	18-24	male	Undergraduate	Law	> 10 Hours	> 10 Hours	Non	Non	Yes
24	18-24	male	Undergraduate	Business	> 10 Hours	> 10 Hours	Limited	Limited	Yes
25	18-24	male	Undergraduate	Business	> 10 Hours	> 10 Hours	Non	Non	Yes
26	18-24	female	Undergraduate	Engineering	> 10 Hours	> 10 Hours	Good	Non	Yes
27	18-24	male	Undergraduate	Computing	> 10 Hours	> 10 Hours	Non	Good	Yes
28	18-24	female	Undergraduate	Art	> 10 Hours	> 10 Hours	Limited	Good	Yes
29	18-24	male	Undergraduate	Law	> 10 Hours	> 10 Hours	Limited	Non	Yes
30	18-24	male	Undergraduate	Art	> 10 Hours	> 10 Hours	Non	Non	Yes

First experiment: tasks performing times experimental condition)

MMC Task Completion Time								
	Easy		Moderate		Difficult			
User	RGN1	RCL1	RGN3	RCL3	RGN5	RCL5	Total Time(Secon	Mean Value
U1	20.41	23.60	25.20	27.00	26.80	31.30	154.31	25.72
U2	19.70	24.65	26.01	24.10	23.90	31.10	149.46	24.91
U3	21.60	23.00	24.12	25.00	24.80	29.00	147.52	24.59
U4	22.40	25.00	22.90	26.50	26.30	29.00	152.10	25.35
U5	21.00	23.00	24.20	25.00	24.80	28.00	146.00	24.33
U6	22.40	26.12	25.64	27.00	26.80	30.00	157.96	26.33
U7	22.80	25.10	26.20	26.30	26.10	29.00	155.50	25.92
U8	18.90	22.90	22.35	24.00	23.80	28.45	140.40	23.40
U9	21.80	24.52	25.00	26.00	25.80	30.00	153.12	25.52
U10	21.20	23.70	24.10	25.00	24.80	29.00	147.80	24.63
U11	21.40	24.00	26.05	27.10	26.90	30.00	155.45	25.91
U12	22.80	25.00	25.10	26.00	25.80	31.00	155.70	25.95
U13	19.00	21.30	20.90	24.90	24.70	27.00	137.80	22.97
U14	21.00	23.00	24.40	25.50	25.30	30.50	149.70	24.95
U15	20.10	24.10	23.90	24.70	24.50	30.00	147.30	24.55
U16	16.40	22.90	21.30	23.20	23.00	27.40	134.20	22.37
U17	22.20	24.10	24.10	24.00	23.80	28.00	146.20	24.37
U18	23.70	25.21	25.10	26.60	26.40	29.00	156.01	26.00
U19	22.00	25.00	24.40	24.70	24.50	28.00	148.60	24.77
U20	23.00	22.60	23.80	24.10	23.90	27.00	144.40	24.07
U21	20.00	23.50	24.10	25.00	24.80	30.20	147.60	24.60
U22	18.00	22.00	21.20	25.00	24.80	27.30	138.30	23.05
U23	22.00	24.10	24.20	25.50	25.30	29.30	150.40	25.07
U24	23.00	25.10	25.00	24.10	23.90	27.20	148.30	24.72
U25	18.90	22.00	22.90	23.20	23.00	29.00	139.00	23.17
U26	20.30	24.00	26.00	26.00	25.80	28.00	150.10	25.02
U27	22.80	24.00	26.20	26.50	26.30	28.20	154.00	25.67
U28	18.70	24.60	22.90	24.00	23.80	26.00	140.00	23.33
U29	22.80	24.00	25.30	26.00	25.80	28.00	151.90	25.32
U30	21.00	24.40	24.60	24.60	24.40	27.00	146.00	24.33

First experiment: tasks performing times (control condition)

NMMC Task Completion Time								
	Easy		Moderate		Difficult			
Users	RGN2	RCL2	;RGN4	RCL4	RGN6	RCL6	Total (Seconds)	Mean Value
U1	21.00	25.60	27.00	29.00	29.70	33.00	165.30	27.55
U2	20.00	26.10	26.60	26.55	28.50	32.00	159.75	26.63
U3	22.00	28.50	27.90	28.00	30.40	31.00	167.80	27.97
U4	22.00	27.10	29.00	28.00	28.50	33.00	167.60	27.93
U5	20.00	29.12	30.10	29.10	31.60	31.00	170.92	28.49
U6	22.30	26.30	28.20	29.00	26.60	31.20	163.60	27.27
U7	21.00	25.70	29.10	28.60	31.60	32.00	168.00	28.00
U8	20.80	28.00	28.90	27.00	27.30	31.00	163.00	27.17
U9	20.00	27.50	24.00	29.50	28.50	30.00	159.50	26.58
U10	22.20	28.00	24.30	28.00	28.90	31.00	162.40	27.07
U11	20.90	26.30	23.00	28.00	26.60	32.00	156.80	26.13
U12	21.00	27.00	26.60	30.00	29.00	31.00	164.60	27.43
U13	22.00	28.20	24.10	27.70	29.40	29.80	161.20	26.87
U14	22.50	26.10	27.20	27.00	25.00	31.00	158.80	26.47
U15	22.50	27.20	28.00	29.00	26.60	32.00	165.30	27.55
U16	19.00	24.00	25.55	28.30	31.10	31.00	158.95	26.49
U17	23.00	26.40	28.00	27.00	27.10	30.00	161.50	26.92
U18	22.10	26.80	25.00	28.20	27.30	31.00	160.40	26.73
U19	21.00	27.70	27.00	27.00	30.00	30.00	162.70	27.12
U20	20.00	29.00	28.10	26.90	28.50	31.00	163.50	27.25
U21	21.00	28.00	29.10	27.00	28.70	32.00	165.80	27.63
U22	19.00	24.00	24.41	28.00	23.40	30.60	149.41	24.90
U23	20.00	25.00	26.60	28.10	30.80	30.00	160.50	26.75
U24	21.00	26.20	25.90	29.00	30.80	30.00	162.90	27.15
U25	19.00	25.00	26.00	28.00	27.50	32.80	158.30	26.38
U26	20.00	24.00	27.00	29.00	26.60	31.10	157.70	26.28
U27	22.00	25.00	25.20	28.00	28.30	30.00	158.50	26.42
U28	19.00	26.00	26.10	27.70	29.80	29.00	157.60	26.27
U29	21.00	27.00	27.00	28.20	31.40	28.00	162.60	27.10
U30	22.00	24.60	26.00	29.20	29.30	30.00	161.10	26.85

First experiment: task performing scores (experimental condition)

MMC Task scores						
	Easy Task		Moderate Task		Difficult Task	
Users	RGN1	RCL1	RGN3	RCL3	RGN5	RCL5
U1	1	1	1	0	1	1
U2	1	1	1	1	1	1
U3	1	1	1	1	0	1
U4	1	1	1	1	1	0
U5	1	1	0	1	1	1
U6	1	1	1	1	1	0
U7	1	1	1	1	1	0
U8	1	1	1	1	1	1
U9	1	1	1	1	1	1
U10	1	1	1	1	0	0
U11	1	1	1	1	1	1
U12	1	0	1	0	1	1
U13	1	1	1	1	1	0
U14	1	1	1	1	1	1
U15	1	1	1	1	0	0
U16	1	1	1	1	1	1
U17	1	1	1	1	0	1
U18	1	1	1	1	1	0
U19	1	1	0	1	1	1
U20	0	1	1	0	1	1
U21	1	1	1	1	1	0
U22	1	1	1	1	1	1
U23	1	1	1	1	0	1
U24	1	1	0	1	1	1
U25	1	1	1	1	0	0
U26	1	1	1	1	1	1
U27	1	1	1	1	0	1
U28	1	1	1	1	1	0
U29	1	1	1	1	0	1
U30	1	0	0	1	1	1

First experiment: task performing scores (control condition).

	NMMC task scores					
Users	Easy Task		Moderate Task		Difficult Task	
	RGN2	RCL2	RGN4	RCL4	RGN6	RCL6
U1	1	1	1	1	1	1
U2	1	1	1	1	0	0
U3	1	1	0	0	1	0
U4	1	1	1	1	1	0
U5	0	1	1	0	0	0
U6	1	0	1	1	0	0
U7	1	1	1	0	1	0
U8	1	1	0	1	1	1
U9	1	1	1	1	0	1
U10	1	1	1	0	0	0
U11	1	1	0	0	1	0
U12	1	1	1	1	0	0
U13	1	1	1	0	1	0
U14	1	1	0	1	1	0
U15	1	0	0	0	0	0
U16	0	1	1	1	0	1
U17	1	1	1	0	0	0
U18	1	1	1	0	1	1
U19	1	1	0	1	0	0
U20	1	1	1	0	0	0
U21	1	0	0	0	0	0
U22	1	1	1	1	0	1
U23	1	1	1	0	1	0
U24	1	1	1	1	0	0
U25	1	1	0	1	1	0
U26	1	1	1	1	1	1
U27	1	1	0	1	0	0
U28	1	1	1	1	1	1
U29	1	0	0	1	0	0
U30	0	1	1	0	0	0

First experiment: users satisfaction scores (control condition)

Users	NMMC Satisfaction Statements																	
	SUS-Statements										Additional - Statements							
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18
U1	1	2	2	3	3	2	2	2	1	2	3	2	3	3	3	2	2	2
U2	3	1	3	2	2	3	3	1	3	1	3	3	4	4	3	2	3	3
U3	1	1	2	1	4	2	2	2	2	1	2	1	3	3	4	2	1	2
U4	3	2	3	1	3	1	3	1	2	1	4	2	4	3	3	1	2	3
U5	2	2	4	2	4	3	4	2	2	2	3	1	3	3	4	1	1	1
U6	1	2	3	2	3	2	3	3	2	1	2	1	4	4	4	1	1	2
U7	2	1	4	2	3	1	4	3	2	2	3	2	3	3	3	1	2	2
U8	1	2	3	2	2	3	3	3	2	2	2	1	4	3	3	1	1	2
U9	2	2	3	2	3	2	3	2	3	2	2	1	4	4	4	1	1	1
U10	3	2	3	1	3	1	3	2	4	1	2	1	3	3	3	1	1	1
U11	1	2	3	2	3	3	3	3	3	2	2	1	4	3	3	1	1	1
U12	2	2	3	2	3	2	3	2	2	3	3	2	3	3	3	1	2	2
U13	2	2	3	2	4	2	3	2	2	2	3	1	4	3	3	1	2	2
U14	3	2	4	3	3	1	4	2	3	1	3	2	3	3	3	1	2	2
U15	2	2	3	1	2	2	3	2	2	2	2	2	4	3	4	1	2	1
U16	3	1	3	1	3	3	3	2	4	1	3	3	4	3	3	1	2	2
U17	3	1	3	1	4	2	3	2	4	1	4	3	4	4	4	1	2	2
U18	2	2	4	2	3	2	3	3	2	2	3	2	4	3	3	1	1	1
U19	3	1	3	3	3	2	3	2	4	1	3	2	3	3	3	1	1	1
U20	2	2	3	3	3	2	3	2	3	3	2	2	4	3	3	2	1	1
U21	2	1	2	2	2	2	2	2	2	2	2	1	3	3	3	3	1	1
U22	1	3	2	1	2	2	2	2	2	1	1	1	3	3	4	3	1	1
U23	2	1	2	1	3	2	2	1	2	1	1	1	3	4	3	3	1	1
U24	2	1	3	3	4	3	3	3	3	2	2	2	3	3	3	3	1	2
U25	3	1	3	3	2	1	3	1	4	1	3	2	4	4	3	2	1	2
U26	2	3	3	1	3	2	3	3	3	1	1	1	3	3	3	3	1	1
U27	3	1	4	2	1	2	4	2	3	2	2	1	3	4	3	3	2	2
U28	3	1	4	3	2	1	4	2	3	1	3	3	4	4	4	4	3	1
U29	3	1	3	2	3	2	3	2	3	1	2	2	3	3	3	3	1	2
U30	2	2	2	3	2	2	2	3	2	2	2	2	4	4	3	3	2	2
Mean	0.54	0.40	0.75	0.49	0.70	0.50	0.74	0.53	0.65	0.39	0.60	0.42	0.87	0.82	0.81	0.45	0.37	0.40

First experiment: users satisfaction scores (experimental condition)

Users	MMC Satisfaction statements																	
	SUS Statements									Additional Statements								
	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S16	S15	S17	S18
U1	4	1	4	2	4	2	3	2	4	2	4	3	3	4	4	4	3	4
U2	3	1	4	1	4	1	3	1	4	1	3	4	4	4	4	4	4	3
U3	4	1	4	1	4	1	4	1	4	1	4	4	4	4	4	4	4	4
U4	4	3	4	3	4	2	4	2	4	2	3	4	4	3	3	4	4	4
U5	4	2	4	4	4	3	4	2	4	3	4	4	4	4	4	3	4	4
U6	3	1	4	2	4	2	4	2	4	1	4	3	4	4	4	3	4	4
U7	3	1	4	1	4	4	3	4	4	1	3	4	3	2	2	3	3	4
U8	4	2	4	2	4	2	4	1	4	2	3	3	4	4	4	3	4	4
U9	4	2	4	3	4	2	4	3	4	2	4	4	4	4	4	4	4	4
U10	3	2	4	1	4	2	3	2	4	2	3	3	3	3	3	2	3	3
U11	3	1	4	1	4	1	3	1	4	1	3	3	3	3	3	3	3	3
U12	3	2	4	2	4	2	4	2	4	2	3	4	4	4	4	3	3	4
U13	4	2	4	2	4	2	3	2	4	4	4	4	4	3	3	3	4	4
U14	3	1	4	1	4	1	3	1	4	2	2	3	2	2	2	2	2	3
U15	4	2	4	4	4	4	4	2	4	3	3	3	3	3	3	3	3	3
U16	4	2	4	1	4	2	4	2	4	2	4	4	4	3	3	3	3	3
U17	4	2	3	3	4	2	4	2	3	4	3	3	3	2	2	4	3	3
U18	3	1	4	1	4	1	3	1	4	1	3	3	3	3	3	2	2	3
U19	3	1	4	1	4	1	4	1	4	4	3	2	2	2	2	2	2	2
U20	3	2	4	2	4	2	4	2	4	2	3	4	4	3	3	3	3	3
U21	3	1	4	1	4	1	3	2	4	1	3	3	2	3	3	2	2	3
U22	2	1	4	1	4	1	3	1	4	1	3	2	3	3	3	3	3	3
U23	3	2	4	2	4	2	3	2	4	2	3	3	4	3	3	3	3	3
U24	3	1	4	1	4	2	3	1	4	2	2	3	2	3	3	3	2	3
U25	4	1	4	1	4	1	4	1	4	1	3	3	3	2	2	3	3	3
U26	3	2	4	3	4	3	3	3	4	3	3	3	3	3	3	3	3	3
U27	4	1	4	1	4	1	4	1	4	2	4	3	3	3	3	4	4	3
U28	4	2	4	1	4	1	4	1	4	1	3	3	3	3	3	3	3	3
U29	3	1	4	2	4	2	3	2	4	2	3	3	3	2	2	2	3	3
U30	4	1	4	1	4	1	3	1	4	1	4	4	4	3	3	3	4	4
Mean	0.85	0.37	0.99	0.43	1.00	0.45	0.87	0.42	0.99	0.48	0.80	0.82	0.82	0.76	0.76	0.75	0.79	0.83

First experiment: user experience measurements data

User experience data			
	Pair of words	NMMC	MMC
PQ	Technical - Human	-1.63	0.77
	Complicated - Simple	1.50	1.20
	Impractical - Practical	1.20	1.63
	cumbersome - Straightforward	1.03	1.30
	Unpredictable - Predictable	0.67	0.93
	Confusing - Clear	0.70	1.57
	Unruly - Manageable	0.23	2.03
HQ-I	Isolating-Connective	-0.73	1.47
	Unprofessional-Professional	-0.97	1.90
	Tacky-Stylish	-1.03	1.87
	Cheap-Premium	-0.93	2.27
	Alienating-Integrating	-0.83	1.67
	Separates me-Brings me closer	-0.57	0.97
	Unpresentable-Presentable	0.00	2.20
HQ-S	Conventional-Inventing	-2.23	2.00
	Unimaginative-Creative	-0.50	2.23
	Cautious-Bold	-0.43	1.40
	Conservative-Innovative	-1.63	2.37
	Dull-Captivating	-0.93	2.13
	Undemanding-Challenging	-0.80	1.47
	Ordinary-Novel	-1.83	1.97
ATT	Unpleasant-Pleasant	-0.03	2.37
	Ugly-Attractive	-0.97	2.33
	Disagreeable-Likable	-0.43	2.10
	Rejecting-Inviting	-0.53	2.07
	Bad - Good	0.70	2.13
	Repelling - Appealing	-0.80	1.90
	Discouraging - Motivation	-0.73	1.80

First experiment: users affective state SAM (valence, arousal and dominance) data

SAM Results						
Conditions	MMC (Experimental condition)			NMMC (Control condition)		
Users	Valence	Arousal	Dominance	Valence	Arousal	Domination
U1	7	8	9	5	6	4
U2	8	8	8	5	5	5
U3	8	7	8	5	6	7
U4	7	8	7	5	8	7
U5	8	7	7	4	6	6
U6	8	6	8	5	4	5
U7	7	6	7	5	5	7
U8	7	7	8	4	4	5
U9	8	8	8	5	6	6
U10	7	5	7	3	6	5
U11	7	6	7	4	3	5
U12	8	8	8	5	6	7
U13	7	7	8	5	6	5
U14	6	5	5	5	5	5
U15	7	7	7	5	6	6
U16	7	7	8	5	7	7
U17	8	6	7	5	5	7
U18	6	6	6	4	7	6
U19	5	5	5	2	6	5
U20	8	7	8	5	4	7
U21	7	7	7	5	6	5
U22	7	6	7	4	6	5
U23	7	7	7	4	4	6
U24	6	6	6	5	6	6
U25	7	6	8	6	6	7
U26	7	5	6	2	6	6
U27	8	7	7	4	6	6
U28	8	8	8	6	6	5
U29	6	5	6	3	7	4
U30	8	8	8	5	7	7

First experiment: users affective state, biofeedback data

Skin conductance and temperature				
	MMC (Experimental condition)		NMMC (Control condition)	
Users	SKT	SC	SKT	SC
U1	32.805	4.084	33.473	3.514
U2	29.315	0.734	29.617	0.809
U3	31.611	6.069	32.076	5.724
U4	31.754	4.345	32.839	3.121
U5	30.600	1.214	32.395	1.075
U6	29.394	2.994	32.689	2.563
U7	32.602	0.228	33.203	0.193
U8	28.928	0.792	26.869	0.768
U9	30.329	0.168	31.115	0.124
U10	30.435	0.714	29.948	0.670
U11	29.668	3.151	30.342	2.505
U12	32.348	2.832	32.899	2.361
U13	30.259	13.351	30.775	11.917
U14	31.820	3.296	31.841	2.187
U15	28.437	1.503	28.584	1.226
U16	33.197	8.397	33.031	4.961
U17	28.227	3.751	28.875	2.339
U18	32.743	3.009	33.384	1.921
U19	32.427	3.741	32.528	2.820
U20	31.510	4.061	31.238	3.642
U21	29.985	0.868	30.228	0.621
U22	24.372	4.418	24.682	3.868
U23	32.636	0.258	32.845	0.233
U24	32.393	0.361	32.258	0.339
U25	28.062	1.067	28.522	1.028
U26	31.464	0.994	31.238	0.943
U27	29.357	1.218	30.702	0.606
U28	29.183	1.596	31.198	0.416
U29	32.351	1.394	32.196	1.381
U30	30.131	2.034	30.740	2.004
Mean	30.611	2.755	31.078	2.196

APPENDIX C

SECOND EXPERIMENT

Second experiment: users' pre-session raw data

	Age Group	Gender	Education Level	Avg. Use of Computer	Avg. Use of Internet	Study Area	HCI Knowledge	Knowledge of the learning topic	Practice of eLearning
U1	25-34	Male	PhD	>10 Hours	>10 Hours	Comp.	Good	Good	Yes
U2	35-44	Male	MSc	>10 Hours	>10 Hours	Comp.	Good	Good	Yes
U3	18-24	Male	UGD	>10 Hours	>10 Hours	Art	Limited	Limited	Yes
U4	25-34	Male	MSc	>10 Hours	>10 Hours	Art	Limited	Limited	Yes
U5	25-34	Male	MSc	>10 Hours	>10 Hours	Art	Limited	Limited	Yes
U6	18-24	Male	UGD	1-5 Hours	1-5 Hours	LANG	Non	Non	No
U7	18-24	Male	UGD	1-5 Hours	>10 Hours	Art	Non	Non	No
U8	18-24	Male	UGD	>10 Hours	>10 Hours	ENG	Non	Non	No
U9	25-34	Female	MSc	6-10 Hours	>10 Hours	HS	Non	Non	No
U10	45+	Male	PhD	6-10 Hours	6-10 Hours	ENG	Non	Non	No
U11	18-24	Male	UGD	>10 Hours	6-10 Hours	ENG	Non	Non	No
U12	35-44	Male	MSc	6-10 Hours	>10 Hours	ENG	Non	Non	No
U13	18-24	Male	UGD	>10 Hours	>10 Hours	ENG	Non	Non	No
U14	18-24	Male	UGD	6-10 Hours	>10 Hours	Art	Non	Non	No
U15	18-24	Female	UGD	>10 Hours	>10 Hours	Art	Non	Non	No
U16	25-34	Female	MSc	>10 Hours	>10 Hours	Art	Non	Non	No
U17	18-24	Male	UGD	>10 Hours	>10 Hours	Law	Non	Non	Yes
U18	18-24	Male	UGD	>10 Hours	>10 Hours	Law	Non	Non	Yes
U19	25-34	Male	MSc	6-10 Hours	>10 Hours	Law	Limited	Limited	Yes
U20	35-44	Male	MSc	>10 Hours	>10 Hours	Law	Limited	Limited	Yes
U21	25-34	Female	MSc	>10 Hours	1-5 Hours	HS	Non	Non	Yes
U22	35-44	Male	MSc	>10 Hours	6-10 Hours	LANG	Non	Non	Yes
U23	35-44	Female	PhD	6-10 Hours	6-10 Hours	BUS	Non	Non	Yes
U24	18-24	Male	UGD	6-10 Hours	6-10 Hours	ENG	Non	Non	Yes
U25	25-34	Male	MSc	6-10 Hours	6-10 Hours	ENG	Non	Non	Yes
U26	18-24	Male	UGD	>10 Hours	>10 Hours	HS	Non	Non	Yes
U27	25-34	Male	MSc	>10 Hours	>10 Hours	ENG	Non	Non	Yes
U28	35-44	Male	PhD	>10 Hours	>10 Hours	ENG	Non	Non	Yes
U29	35-44	Male	PhD	>10 Hours	>10 Hours	HS	Non	Non	Yes
U30	18-24	Female	UGD	>10 Hours	>10 Hours	HS	Non	Non	Yes

Second experiment: auditory condition (experiment and task performance) time

			Auditory Condition										
			Recall Tasks				Recognition Tasks						
User	Auditory Feedback Time	Lesson Learning Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average Recognition Time	Total Task Time	Total Experiment Time	average task time
U1	146.88	320.00	22.19	23.28	24.20	23.22	20.22	21.64	22.44	21.43	157.19	624.07	123.98
U2	120.96	335.00	22.44	20.18	22.25	21.62	18.46	18.31	20.15	18.97	143.41	599.37	110.19
U3	216.86	345.00	21.14	24.59	23.19	22.97	17.54	23.34	21.08	20.65	153.85	715.71	136.59
U4	103.68	315.00	18.58	19.56	21.46	19.87	18.13	18.37	22.56	19.69	138.54	557.22	104.58
U5	138.24	323.00	19.54	22.02	23.50	21.69	17.30	19.49	23.58	20.13	147.12	608.36	112.67
U6	121.82	314.00	22.23	25.56	26.12	24.63	20.33	23.56	24.55	22.81	166.98	602.80	116.19
U7	155.52	330.00	24.45	24.31	26.09	24.95	19.02	21.51	23.48	21.34	163.81	649.33	151.59
U8	170.88	345.00	22.27	24.59	25.47	24.11	20.26	22.19	22.12	21.52	161.01	676.89	127.49
U9	122.69	340.00	19.29	23.38	25.16	22.61	19.13	20.25	22.51	20.63	152.33	615.02	111.59
U10	190.08	300.00	21.39	22.12	25.03	22.84	19.06	20.34	23.00	20.80	153.78	643.86	116.86
U11	164.16	350.00	22.54	22.27	24.22	23.01	20.18	21.49	24.01	21.89	157.72	671.88	120.26
U12	156.38	300.00	20.18	21.34	22.48	21.33	17.23	19.03	21.17	19.14	142.75	599.14	111.83
U13	147.00	320.00	21.03	23.06	24.32	22.80	20.40	21.05	22.16	21.20	154.82	621.82	117.70
U14	138.24	350.00	19.49	21.22	23.01	21.24	17.19	21.30	22.04	20.17	145.50	633.74	118.17
U15	142.56	335.00	20.39	21.41	22.44	21.41	18.28	19.02	21.44	19.58	144.38	621.94	117.00
U16	122.69	345.00	24.09	25.17	26.16	25.14	18.41	20.14	23.17	20.58	162.29	629.98	117.82
U17	104.54	340.00	18.14	23.59	24.25	21.99	17.25	19.46	23.59	20.10	148.27	592.82	111.50
U18	132.45	332.00	18.45	21.09	23.46	21.00	16.13	18.09	20.32	18.18	138.54	602.99	109.89
U19	105.43	290.00	17.55	22.18	23.30	21.01	15.38	20.18	21.28	18.95	140.88	536.31	100.71
U20	157.42	300.00	17.15	24.48	24.18	21.94	16.13	18.40	20.50	18.34	142.78	600.20	109.90
U21	147.83	328.00	19.08	22.50	26.25	22.61	16.42	20.27	20.09	18.93	147.22	623.05	114.71
U22	147.00	331.00	20.19	22.18	22.35	21.57	19.54	20.14	22.26	20.65	148.23	626.23	116.90
U23	147.00	322.00	18.29	23.18	25.60	22.35	17.25	18.26	23.13	19.55	148.06	617.06	113.81
U24	160.70	301.00	20.57	22.00	24.31	22.29	18.08	21.49	22.27	20.62	151.02	612.72	114.69
U25	172.80	311.00	19.28	21.42	25.44	22.05	20.26	19.58	23.43	21.09	151.44	635.24	119.44
U26	164.16	290.00	21.32	23.13	24.23	22.90	20.36	21.12	24.12	21.87	157.18	611.34	116.56
U27	171.88	360.00	19.17	21.21	24.24	21.54	18.57	19.13	20.36	19.35	144.22	676.10	149.33
U28	139.97	335.00	22.46	23.30	25.12	23.63	17.07	21.30	24.44	20.94	157.32	632.29	120.00
U29	170.21	380.00	21.34	22.34	26.20	23.29	19.11	20.49	22.01	20.54	154.77	704.98	138.88
U30	142.56	311.00	19.28	21.35	21.52	20.71	15.27	19.40	23.32	19.33	140.85	594.41	109.18

Second experiment: avatar condition (experiment and task performance) time

			Avatar Task Time										
			Recall Tasks				Recognition Tasks						
User	Auditory Feedback Time	Lesson Learning Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average RecognitionTime	Total Task Time	Total Experiment Time	average task time
U1	140	295	16.67	17.24	18.32	17.41	15.27	19.25	21.50	18.67	125.65	560.65	111.20
U2	120	300	16.88	18.91	19.83	18.54	17.49	18.11	20.34	18.65	130.11	550.11	110.48
U3	201	290	19.32	21.88	22.46	21.22	17.50	18.33	20.79	18.87	141.50	632.50	124.39
U4	109	285	15.71	17.80	18.37	17.29	14.45	19.35	20.24	18.01	123.21	517.21	104.25
U5	153	290	18.30	17.38	19.04	18.24	17.03	18.04	20.92	18.66	128.97	571.97	113.40
U6	132	283	18.88	20.72	21.02	20.21	16.53	17.05	19.20	17.59	133.61	548.61	110.40
U7	150	292	18.13	19.82	21.33	19.76	18.05	18.35	19.49	18.63	134.94	576.94	115.17
U8	160	295	16.82	17.02	18.40	17.41	17.71	20.03	21.43	19.72	128.83	583.83	115.57
U9	133	293	17.86	19.00	21.90	19.59	15.42	18.45	19.52	17.80	131.73	557.73	111.46
U10	166	293	18.55	20.14	21.15	19.94	16.20	17.73	21.59	18.51	135.30	594.30	117.65
U11	161	287	16.73	19.37	18.80	18.30	15.29	16.00	17.36	16.22	121.85	569.85	110.70
U12	152	290	17.06	17.50	21.54	18.70	16.29	17.25	18.57	17.37	126.93	568.93	112.01
U13	140	280	17.58	19.46	20.86	19.30	14.71	17.69	19.55	17.32	129.15	549.15	109.55
U14	135	284	18.05	20.41	20.55	19.67	16.51	17.33	20.33	18.06	132.86	551.86	110.94
U15	140	302	18.94	20.59	22.20	20.57	16.50	18.29	20.14	18.31	137.23	579.23	115.75
U16	110	291	18.91	19.94	20.60	19.82	18.65	19.67	20.67	19.66	138.26	539.26	110.86
U17	120	294	18.64	19.52	21.82	19.99	17.14	18.92	19.56	18.54	135.60	549.60	111.34
U18	130	292	17.53	18.22	20.33	18.69	15.86	19.02	21.73	18.87	131.40	553.40	111.28
U19	105	285	18.96	19.01	19.28	19.09	17.42	19.19	21.39	19.33	134.34	524.34	107.87
U20	155	289	20.60	21.00	23.15	21.59	15.22	16.51	18.11	16.61	136.19	580.19	114.92
U21	142	297	18.66	21.35	22.45	20.82	16.23	17.05	20.90	18.06	137.48	576.48	115.29
U22	141	300	19.74	21.36	23.55	21.55	15.64	17.45	19.61	17.57	138.91	579.91	115.80
U23	140	290	19.82	19.35	22.99	20.72	16.10	16.56	21.37	18.01	136.90	566.90	113.79
U24	160	295	18.05	19.13	23.48	20.22	18.97	19.48	20.80	19.75	140.13	595.13	119.21
U25	164	279	17.77	18.44	19.90	18.70	15.14	18.75	21.42	18.44	130.13	573.13	113.67
U26	160	284	18.54	19.00	20.28	19.27	15.38	16.96	18.55	16.97	127.99	571.99	112.45
U27	168	290	17.01	18.31	22.98	19.43	16.89	17.17	18.92	17.66	130.70	588.70	115.64
U28	141	302	17.29	18.78	23.47	19.84	16.56	19.08	20.64	18.76	135.65	578.65	115.60
U29	163	300	18.04	21.25	21.32	20.20	16.39	18.06	21.04	18.50	136.30	599.30	118.54
U30	140	290	17.69	18.46	20.77	18.97	15.14	18.29	19.19	17.54	128.53	558.53	110.89

Second experiment: speech condition (experiment and task performance) time

			Speech Condition										
			Recall				Recognition						
User	Auditory Feedback Time	Lesson Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average RecognitionTime	Total Task Time	Total Experiment Time	average task time
U1	165.60	320.00	19.23	21.68	24.05	21.65	19.02	19.43	23.03	20.49	148.09	633.68	126.49
U2	138.68	268.00	20.54	22.46	23.64	22.22	20.06	20.74	21.55	20.78	151.22	557.89	116.35
U3	232.58	298.00	18.70	19.70	21.40	19.93	18.15	19.94	20.85	19.65	138.68	669.26	129.49
U4	179.00	305.00	17.67	20.14	21.66	19.83	20.05	20.34	20.34	20.24	140.03	624.03	123.55
U5	153.30	285.00	17.57	20.26	24.59	20.81	17.78	20.73	21.33	19.95	143.06	581.37	117.86
U6	139.89	310.00	20.40	21.66	24.99	22.35	17.81	20.48	24.05	20.78	151.75	601.63	122.69
U7	175.58	262.00	20.22	21.75	25.82	22.60	16.98	18.40	21.71	19.03	147.47	585.06	118.75
U8	188.94	297.00	18.58	20.94	21.40	20.31	17.76	19.46	21.08	19.43	139.52	625.46	123.29
U9	137.75	271.00	18.29	22.15	22.55	21.00	17.06	20.08	23.04	20.06	144.17	552.92	114.05
U10	207.14	285.00	20.20	20.86	22.61	21.22	16.81	19.65	23.75	20.07	145.10	637.24	126.26
U11	182.22	261.00	19.22	20.60	22.11	20.64	16.80	20.37	21.39	19.52	141.13	584.36	117.74
U12	176.45	288.00	18.89	22.72	23.40	21.67	18.75	18.94	20.31	19.33	144.67	609.12	121.83
U13	168.00	274.00	21.36	21.78	22.77	21.97	18.59	19.66	22.03	20.10	148.17	590.17	120.10
U14	153.30	301.00	17.66	19.64	21.87	19.72	16.46	20.34	23.05	19.95	138.75	593.05	118.76
U15	159.49	319.00	19.29	21.70	23.74	21.58	17.73	18.99	20.31	19.01	143.34	621.82	123.25
U16	143.61	273.00	21.28	20.03	22.70	21.33	18.25	20.36	21.51	20.04	145.45	562.07	115.57
U17	119.47	303.00	17.17	21.42	24.43	21.00	17.62	19.34	22.77	19.91	143.75	566.22	115.80
U18	144.90	297.00	15.16	21.56	22.68	19.80	17.64	18.84	20.27	18.92	135.95	577.84	115.61
U19	160.00	280.00	16.65	21.43	24.53	20.87	19.51	19.28	22.06	20.29	144.35	584.35	118.67
U20	173.08	305.00	17.78	22.55	21.21	20.52	17.21	18.63	22.33	19.39	140.24	618.31	122.38
U21	165.48	272.00	19.00	22.88	21.97	21.28	17.82	20.76	20.45	19.68	144.15	581.64	117.97
U22	160.88	256.00	23.50	20.50	21.49	21.83	18.30	18.38	20.67	19.12	144.67	561.55	114.93
U23	164.00	261.00	18.14	20.66	22.81	20.54	16.94	20.65	23.38	20.32	143.12	568.12	116.15
U24	179.36	295.00	20.54	21.26	23.76	21.86	17.21	20.91	21.29	19.80	146.83	621.19	124.16
U25	194.45	305.00	22.92	22.09	25.95	23.66	19.81	20.85	21.85	20.84	157.14	656.60	131.54
U26	183.81	260.00	21.36	22.23	23.24	22.28	19.86	18.43	22.78	20.36	150.17	593.99	121.12
U27	194.53	321.00	21.53	21.87	23.45	22.28	18.61	20.99	23.14	20.92	151.88	667.42	132.18
U28	162.62	299.00	20.47	21.03	24.70	22.07	18.85	21.19	23.01	21.02	151.31	612.93	124.34
U29	191.86	280.00	22.58	23.33	25.79	23.90	16.65	21.05	20.39	19.36	153.67	625.53	125.79
U30	160.21	285.00	19.02	21.76	24.30	21.69	16.54	18.57	20.39	18.50	142.26	587.47	117.92

Second experiment: earcons condition (experiment and task performance) time

			Earcons Condition										
			Recall Tasks				Recognition Tasks						
User	Auditory Feedback Time	Lesson Learning Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average RecognitionTime	Total Task Time	Total Experiment Time	average task time
U1	140	352	22.35	21.35	25.04	22.91	19.33	20.63	23.79	21.25	155.41	647.41	130.11
U2	120	352	22.25	22.37	23.46	22.69	20.67	21.58	20.20	20.82	153.21	625.21	126.34
U3	201	395	20.46	24.32	24.02	22.93	22.63	19.36	24.04	22.01	157.78	753.78	146.08
U4	109	310	19.04	22.63	25.55	22.40	16.18	19.64	21.04	18.95	146.47	565.47	115.74
U5	153	351	20.13	23.21	25.37	22.90	22.10	20.93	22.19	21.74	156.83	660.83	132.50
U6	132	335	23.09	22.42	23.86	23.12	17.69	18.14	23.14	19.66	151.46	618.46	124.53
U7	150	341	22.41	24.22	23.82	23.48	21.17	22.28	22.17	21.88	159.57	650.57	131.59
U8	160	340	23.15	23.02	23.03	23.07	21.03	21.30	21.60	21.31	156.19	656.19	131.53
U9	133	335	20.17	21.92	24.98	22.36	17.79	21.76	23.92	21.16	152.90	620.90	125.83
U10	166	350	20.65	25.00	25.68	23.78	18.75	22.37	23.77	21.63	159.99	675.99	135.18
U11	161	390	21.44	22.61	23.75	22.60	19.54	19.68	20.03	19.75	149.65	700.65	135.99
U12	152	335	23.54	24.70	24.04	24.09	17.08	22.43	22.45	20.65	158.33	645.33	130.05
U13	140	345	22.10	24.12	26.17	24.13	16.54	21.49	20.06	19.36	154.61	639.61	127.97
U14	135	321	20.95	22.03	23.24	22.07	16.70	18.84	23.37	19.64	147.21	603.21	121.58
U15	140	324	20.33	23.46	24.60	22.80	18.32	22.63	21.35	20.77	153.49	617.49	125.27
U16	110	331	21.54	23.89	24.61	23.35	18.77	22.76	22.00	21.18	156.93	597.93	123.27
U17	120	345	20.47	23.50	25.73	23.23	19.65	20.22	24.62	21.50	157.41	622.41	127.01
U18	130	333	19.02	24.41	24.22	22.55	16.71	22.52	23.86	21.03	153.28	616.28	125.18
U19	105	321	18.50	22.05	26.20	22.25	17.42	22.02	22.42	20.62	150.85	577.28	118.98
U20	155	323	18.13	24.83	25.27	22.75	18.32	20.17	21.27	19.92	150.74	628.74	125.99
U21	142	312	20.39	21.10	25.80	22.43	18.05	20.79	20.11	19.65	148.67	602.67	121.77
U22	141	335	23.64	21.50	26.04	23.73	17.72	20.21	20.51	19.48	153.35	629.35	126.33
U23	140	333	20.18	21.84	25.02	22.35	17.39	18.89	20.78	19.02	146.45	619.45	123.47
U24	160	355	20.80	22.73	23.84	22.46	15.86	19.81	24.31	19.99	149.81	664.81	131.01
U25	164	331	22.49	23.42	23.05	22.99	15.97	22.14	22.64	20.25	152.70	647.70	129.20
U26	160	330	20.18	24.53	26.70	23.81	14.18	19.38	23.42	18.99	152.21	642.21	127.74
U27	168	333	22.23	24.89	25.48	24.20	17.44	19.77	21.38	19.53	155.40	656.40	130.59
U28	141	324	21.09	23.60	26.80	23.83	18.65	19.25	23.82	20.57	157.04	622.04	126.46
U29	163	352	23.52	23.44	24.55	23.84	20.45	20.19	21.61	20.75	157.60	672.60	133.86
U30	140	331	20.27	24.62	25.76	23.55	17.15	18.97	20.56	18.89	150.88	621.88	124.55

Second experiment: music condition (experiment and task performance) time

			Classic Music Condition										
			Recall Tasks				Recognition Tasks						
User	Auditory Feedback Time	Lesson Learning Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average RecognitionTime	Total Task Time	Total Experiment Time	average task time
U1	125.88	332.25	19.29	19.16	22.48	20.31	18.15	18.61	20.38	19.04	138.38	596.51	118.77
U2	110.83	333.04	20.17	21.87	22.24	21.43	18.61	19.81	22.80	20.41	146.94	590.81	120.12
U3	129.04	324.33	20.91	21.87	22.91	21.90	20.14	18.80	19.36	19.43	145.89	599.26	120.68
U4	98.96	335.42	20.04	22.67	24.38	22.36	17.82	19.10	20.74	19.22	147.11	581.48	118.26
U5	138.54	328.29	17.90	21.93	21.04	20.29	17.57	17.92	22.75	19.42	139.41	606.24	120.51
U6	117.96	344.13	17.81	20.74	23.18	20.58	14.99	16.97	22.07	18.01	136.34	598.43	118.20
U7	135.38	327.50	19.19	20.99	24.36	21.51	17.14	18.29	21.84	19.09	143.33	606.20	121.06
U8	127.46	321.17	18.09	21.08	24.14	21.10	16.91	18.82	22.31	19.34	142.45	591.08	118.86
U9	118.75	337.79	19.59	20.93	23.04	21.19	19.15	20.47	21.25	20.29	145.62	602.16	121.45
U10	138.54	339.38	19.07	20.89	24.94	21.63	16.21	18.58	19.77	18.19	141.10	619.01	122.07
U11	142.50	303.75	18.50	20.99	21.05	20.18	16.59	18.29	19.65	18.18	135.25	581.50	115.66
U12	133.79	311.67	19.31	20.26	24.97	21.51	16.15	17.78	22.45	18.79	142.42	587.88	118.14
U13	123.50	335.42	18.45	21.56	23.91	21.31	15.38	17.19	19.82	17.47	137.63	596.55	117.91
U14	119.54	324.33	15.66	20.89	23.57	20.04	17.63	18.57	21.15	19.11	137.50	581.38	116.48
U15	125.08	303.75	19.18	20.15	24.32	21.21	16.46	19.83	21.61	19.30	142.75	571.58	116.11
U16	101.33	318.79	19.07	19.86	23.87	20.93	15.87	19.58	19.71	18.38	138.89	559.02	113.20
U17	110.04	311.67	17.81	22.32	22.32	20.82	17.25	19.30	22.71	19.75	142.53	564.24	115.23
U18	115.58	297.42	15.60	21.30	22.12	19.67	15.01	17.56	19.68	17.42	130.95	543.95	109.18
U19	96.92	309.29	16.81	21.74	22.59	20.38	16.63	18.75	20.95	18.78	137.86	544.07	111.06
U20	136.96	315.63	19.33	21.87	22.55	21.25	15.43	16.80	22.59	18.27	139.82	592.41	118.08
U21	125.08	335.42	17.73	21.05	21.89	20.22	15.80	16.28	22.59	18.22	135.57	596.07	117.82
U22	125.08	297.42	17.43	21.48	21.73	20.21	16.45	19.41	20.03	18.63	136.74	559.24	112.96
U23	124.29	307.71	19.04	22.32	24.69	22.02	17.26	18.47	21.10	18.94	144.90	576.90	117.08
U24	138.54	296.63	19.24	19.79	21.74	20.26	16.55	19.32	22.13	19.33	139.02	574.19	115.83
U25	141.71	328.29	18.85	19.38	22.36	20.19	18.43	19.34	20.03	19.26	138.57	608.57	120.63
U26	137.75	309.29	18.03	21.91	21.85	20.60	16.75	19.38	22.22	19.45	140.75	587.79	118.14
U27	114.79	319.58	20.18	21.10	22.10	21.13	16.88	18.07	22.08	19.01	141.53	575.91	116.37
U28	125.88	328.29	16.81	19.29	22.93	19.68	20.41	19.01	21.46	20.29	139.59	593.76	119.17
U29	144.88	318.79	19.83	21.05	24.64	21.84	17.53	19.81	21.29	19.54	145.99	609.66	122.24
U30	124.29	336.21	18.27	21.27	24.68	21.41	16.19	19.51	22.57	19.42	143.89	604.39	121.05

Second experiment: nmmc condition (experiment and task performance) time

			NMMC Condition										
			Recall				Recognition						
User	Auditory Feedback Time	Lesson Time	Easy	Moderate	Difficult	Average Recall Time	Easy	Moderate	Difficult	Average Recognition Time	Total Task Time	Total Experiment Time	average task time
U1	174	380.00	22.65	23.93	25.24	23.94	20.28	21.38	22.21	21.29	160	714	140.34
U2	169	315.00	21.23	22.84	23.12	22.40	18.85	19.01	22.06	19.97	150	634	126.47
U3	129	345.00	22.43	22.27	24.18	22.96	19.52	23.41	23.11	22.02	158	632	128.69
U4	121	295.00	19.49	22.03	23.68	21.73	21.87	22.95	23.70	22.84	155	571	120.00
U5	129	403.00	21.05	22.10	24.77	22.64	21.32	23.77	25.32	23.47	161	693	138.63
U6	120	394.00	22.72	24.84	26.92	24.83	20.55	23.62	25.18	23.12	169	683	138.38
U7	216	410.00	22.11	24.64	24.31	23.69	21.54	22.89	25.40	23.28	165	791	153.13
U8	231	355.00	21.50	22.67	25.14	23.10	18.40	19.93	23.78	20.71	155	741	143.00
U9	144	390.00	23.69	23.56	24.14	23.80	21.45	22.35	22.90	22.23	162	696	138.65
U10	134	380.00	23.03	24.90	24.22	24.05	19.90	20.15	23.74	21.26	160	674	134.73
U11	184	330.00	22.31	25.98	25.94	24.74	21.84	23.07	23.00	22.64	167	681	137.59
U12	134	380.00	21.48	23.67	23.38	22.84	21.05	22.69	25.93	23.22	161	675	135.97
U13	154	400.00	21.21	25.20	25.71	24.04	21.45	22.31	23.34	22.36	163	717	142.00
U14	184	330.00	22.03	23.19	23.41	22.88	20.32	22.68	23.12	22.04	158	672	134.33
U15	169	315.00	21.80	22.32	23.40	22.51	18.84	20.08	22.46	20.46	151	635	127.31
U16	179	325.00	24.25	25.31	24.77	24.78	19.60	19.46	23.06	20.71	161	665	133.44
U17	174	320.00	23.54	24.22	24.44	24.07	20.44	21.85	23.92	22.07	162	656	133.04
U18	166	312.00	21.25	23.36	25.79	23.47	19.82	19.68	23.36	20.96	157	635	128.39
U19	124	320.00	20.89	22.39	25.30	22.86	20.69	21.82	22.12	21.54	156	600	123.60
U20	134	380.00	22.43	24.02	24.33	23.59	21.25	23.87	24.11	23.08	164	678	136.73
U21	162	357.00	22.20	24.69	26.83	24.57	21.54	22.88	23.36	22.59	166	685	138.01
U22	165	351.00	20.91	21.84	24.86	22.54	18.55	20.41	24.47	21.14	154	670	132.90
U23	156	341.00	20.89	22.17	22.38	21.81	22.46	22.55	23.95	22.98	156	653	131.88
U24	135	340.00	18.80	22.84	22.79	21.48	18.29	21.44	23.00	20.91	149	624	125.34
U25	145	321.00	22.15	24.22	26.94	24.44	18.44	20.09	22.87	20.47	159	625	127.23
U26	124	325.00	22.74	25.73	24.21	24.23	18.73	19.95	22.75	20.48	158	607	124.54
U27	216	390.00	21.76	23.34	25.47	23.53	19.97	23.39	24.21	22.52	162	768	148.99
U28	141	340.00	22.51	23.38	25.55	23.81	19.72	20.48	23.14	21.11	159	640	129.50
U29	246	350.00	23.10	24.64	24.39	24.04	18.53	19.26	23.89	20.56	158	754	145.43
U30	158	391.00	22.53	23.07	25.89	23.83	18.39	21.64	23.99	21.34	159	708	139.56

Second experiment: auditory condition recall and recognition task scores

	Auditory Condition						
	Recall			Recognition			
User	Easy	Moderate	Difficult	Easy	Moderate	Difficult	Overall Score
U1	1	1	1	1	1	0	0.83
U2	1	1	1	1	0	1	0.83
U3	1	1	0	0	0	0	0.33
U4	0	1	1	1	1	0	0.67
U5	1	0	1	1	1	1	0.83
U6	0	1	0	1	1	1	0.67
U7	1	0	1	0	0	0	0.33
U8	0	0	0	1	1	0	0.33
U9	1	1	1	1	1	1	1.00
U10	1	0	1	0	1	0	0.50
U11	1	1	0	1	0	1	0.67
U12	1	0	1	0	1	1	0.67
U13	1	1	1	1	1	0	0.83
U14	0	0	0	0	0	0	0.00
U15	0	1	1	1	0	0	0.50
U16	1	0	0	1	0	1	0.50
U17	0	1	1	1	1	0	0.67
U18	1	0	0	0	0	1	0.33
U19	1	1	0	1	1	0	0.67
U20	1	0	1	1	0	1	0.67
U21	1	0	0	0	1	0	0.33
U22	1	0	0	1	0	0	0.33
U23	0	1	1	0	0	1	0.50
U24	0	1	0	1	1	0	0.50
U25	1	0	0	1	1	0	0.50
U26	1	1	1	1	1	1	1.00
U27	0	0	0	1	0	0	0.17
U28	1	1	0	0	0	1	0.50
U29	1	1	0	1	1	0	0.67
U30	0	0	1	1	1	0	0.50

Second experiment: avatar condition recall and recognition task score

	Avatar Condition Task Scores						
	Recall			Recognition			
User	Easy	Moderate	Difficult	Easy	Moderate	Difficult	Overall Score
U1	1	1	0	1	1	1	0.83
U2	1	1	1	1	1	0	0.83
U3	1	0	1	1	1	1	0.83
U4	1	1	0	1	1	1	0.83
U5	1	0	1	1	1	0	0.67
U6	1	0	1	1	1	1	0.83
U7	1	1	1	1	1	0	0.83
U8	1	0	1	1	1	1	0.83
U9	1	1	0	1	0	1	0.67
U10	1	1	1	0	1	1	0.83
U11	1	1	1	1	1	1	1.00
U12	1	1	1	1	1	0	0.83
U13	1	0	1	1	1	1	0.83
U14	1	1	1	1	1	1	1.00
U15	1	1	1	1	1	0	0.83
U16	0	1	1	1	1	1	0.83
U17	1	1	1	0	0	0	0.50
U18	1	1	1	1	1	1	1.00
U19	1	1	0	1	1	1	0.83
U20	1	1	0	1	1	0	0.67
U21	1	1	1	1	1	1	1.00
U22	0	1	0	1	1	0	0.50
U23	1	1	1	1	0	1	0.83
U24	1	1	1	1	1	1	1.00
U25	1	1	1	1	0	0	0.67
U26	1	1	1	1	1	1	1.00
U27	1	1	0	1	1	1	0.83
U28	0	1	1	1	1	0	0.67
U29	1	1	1	1	0	1	0.83
U30	1	1	1	1	1	1	1.00

Second experiment: earcons condition (recall and recognition task scores)

	Earcons Condition Task Scores						
	Recall			Recognition			
User	Easy	Moderate	Difficult	Easy	Moderate	Difficult	Overall Score
U1	1	1	0	0	0	1	0.50
U2	0	0	0	1	1	0	0.33
U3	0	0	1	1	0	0	0.33
U4	0	1	0	0	0	1	0.33
U5	0	0	0	1	1	0	0.33
U6	0	1	0	0	1	1	0.50
U7	1	0	0	0	0	0	0.17
U8	0	0	1	1	0	0	0.33
U9	1	1	0	1	0	0	0.50
U10	0	0	0	1	1	0	0.33
U11	1	0	0	1	1	1	0.67
U12	0	0	1	0	1	1	0.50
U13	0	1	0	1	1	0	0.50
U14	0	1	1	0	0	1	0.50
U15	1	1	0	1	1	0	0.67
U16	1	0	0	1	0	0	0.33
U17	1	0	0	1	0	0	0.33
U18	1	1	0	1	0	0	0.50
U19	1	0	1	0	1	0	0.50
U20	0	1	0	0	1	0	0.33
U21	1	0	0	0	0	1	0.33
U22	0	0	0	0	0	0	0.00
U23	1	1	0	1	0	0	0.50
U24	0	0	1	0	1	0	0.33
U25	1	1	0	1	1	0	0.67
U26	1	1	0	0	0	0	0.33
U27	1	0	0	0	1	1	0.50
U28	0	0	0	1	0	0	0.17
U29	1	0	1	1	0	0	0.50
U30	1	1	0	1	1	1	0.83

Second experiment: classic music condition recall and recognition task score

	Classic Music Condition Scores						
	Recall			Recognition			
User	Easy	Moderate	Difficult	Easy	Moderate	Difficult	Overall Score
U1	1	1	0	1	1	1	0.83
U2	1	1	1	1	1	0	0.83
U3	0	1	1	1	1	1	0.83
U4	0	0	0	1	1	1	0.50
U5	1	1	1	1	0	1	0.83
U6	1	1	0	1	1	1	0.83
U7	1	1	1	1	1	1	1.00
U8	0	1	0	1	1	1	0.67
U9	1	0	1	1	1	0	0.67
U10	1	0	0	0	1	0	0.33
U11	1	1	1	0	1	1	0.83
U12	1	0	1	1	1	1	0.83
U13	1	1	1	1	0	0	0.67
U14	1	1	1	0	1	1	0.83
U15	1	1	1	1	1	1	1.00
U16	1	1	0	1	0	1	0.67
U17	1	1	1	1	1	0	0.83
U18	1	1	1	1	0	1	0.83
U19	0	0	1	1	1	0	0.50
U20	1	1	1	1	0	1	0.83
U21	1	0	0	1	1	1	0.67
U22	0	1	1	1	1	1	0.83
U23	1	1	1	0	1	0	0.67
U24	1	1	0	1	0	1	0.67
U25	1	1	1	1	1	0	0.83
U26	1	0	0	1	1	1	0.67
U27	1	0	1	1	1	0	0.67
U28	1	1	0	1	0	1	0.67
U29	1	1	1	1	1	1	1.00
U30	0	1	1	1	1	0	0.67

Second experiment: speech condition recall and recognition task scores

	Speech Condition Task Scores						
	Recall			Recognition			
User	Easy	Moderate	Difficult	Easy	Moderate	Difficult	Overall Score
U1	1	1	1	1	1	0	0.83
U2	0	1	1	1	0	1	0.67
U3	1	1	1	1	0	1	0.83
U4	1	0	1	1	1	0	0.67
U5	1	1	0	1	0	1	0.67
U6	0	1	1	1	1	0	0.67
U7	1	1	0	0	1	1	0.67
U8	1	0	1	1	1	1	0.83
U9	1	1	0	1	1	1	0.83
U10	1	0	1	1	1	0	0.67
U11	1	1	0	1	0	1	0.67
U12	1	1	1	0	1	1	0.83
U13	1	1	0	1	1	1	0.83
U14	1	1	0	1	1	1	0.83
U15	1	1	1	1	1	1	1.00
U16	1	0	1	1	1	0	0.67
U17	1	1	1	1	1	1	1.00
U18	1	1	1	1	1	0	0.83
U19	1	1	0	1	1	1	0.83
U20	1	1	1	1	1	1	1.00
U21	0	1	1	1	0	1	0.67
U22	1	0	1	0	1	1	0.67
U23	0	1	1	1	1	1	0.83
U24	1	1	1	1	1	0	0.83
U25	1	1	0	0	1	1	0.67
U26	1	0	1	1	0	1	0.67
U27	1	1	1	0	1	1	0.83
U28	1	0	1	1	1	0	0.67
U29	0	1	1	1	1	1	0.83
U30	1	1	1	1	1	0	0.83

Second experiment: nmmc condition recall and recognition task scores

	NMMC Condition Task Scores								
	Recall				Recognition			
User	Easy	Moderate	Difficult	Recall-Mean	Easy	Moderate	Difficult	Recognition-Mean
U1	1	1	1	1.00	0	1	0	0.33
U2	1	1	0	0.67	1	0	0	0.33
U3	1	1	0	0.67	1	1	1	1.00
U4	0	0	1	0.33	1	0	1	0.67
U5	1	1	1	1.00	1	1	1	1.00
U6	1	1	0	0.67	1	1	0	0.67
U7	1	1	1	1.00	1	0	1	0.67
U8	0	0	1	0.33	1	1	0	0.67
U9	0	1	1	0.67	1	1	1	1.00
U10	1	1	1	1.00	1	0	1	0.67
U11	1	1	1	1.00	1	1	0	0.67
U12	1	1	0	0.67	1	1	1	1.00
U13	0	0	1	0.33	0	1	0	0.33
U14	1	1	0	0.67	0	1	0	0.33
U15	0	0	1	0.33	1	1	1	1.00
U16	1	1	0	0.67	1	1	0	0.67
U17	0	0	1	0.33	1	1	1	1.00
U18	1	1	0	0.67	1	0	1	0.67
U19	1	1	0	0.67	1	1	1	1.00
U20	1	0	1	0.67	1	1	0	0.67
U21	1	0	1	0.67	0	0	1	0.33
U22	1	1	0	0.67	0	1	1	0.67
U23	1	0	1	0.67	1	1	0	0.67
U24	1	1	0	0.67	1	1	0	0.67
U25	1	1	0	0.67	1	0	1	0.67
U26	1	1	1	1.00	1	1	0	0.67
U27	1	1	0	0.67	1	0	1	0.67
U28	0	0	0	0.00	1	1	0	0.67
U29	1	1	1	1.00	0	1	1	0.67
U30	1	0	0	0.33	1	0	1	0.67

Second experiment: auditory condition satisfaction statements

Auditory Icons Satisfaction Statements								
User	S1	S2	S3	S4	S5	S6	S7	S8
U1	2	2	1	3	2	3	1	3
U2	2	2	1	2	2	3	1	2
U3	2	3	2	2	2	2	2	2
U4	3	3	3	3	2	3	1	3
U5	3	2	2	2	2	2	2	2
U6	2	2	1	2	2	3	2	3
U7	3	2	2	3	2	3	1	3
U8	3	2	2	2	2	2	1	2
U9	3	3	3	3	2	3	2	1
U10	2	3	2	2	2	2	1	3
U11	3	2	1	3	2	3	2	2
U12	3	3	2	2	2	3	2	3
U13	3	3	3	2	2	3	2	3
U14	3	3	2	2	2	3	1	2
U15	3	3	3	2	3	3	3	2
U16	3	2	2	2	2	3	1	3
U17	3	3	2	2	2	3	2	2
U18	3	2	3	2	2	3	2	2
U19	2	2	1	2	2	2	2	2
U20	3	3	2	2	2	3	2	2
U21	2	2	1	2	2	3	1	3
U22	3	3	3	2	1	3	1	2
U23	3	2	2	2	2	3	1	1
U24	3	3	3	2	2	3	1	3
U25	2	2	1	2	2	3	1	2
U26	3	2	3	2	2	2	2	2
U27	3	3	2	2	2	3	2	3
U28	2	3	2	2	3	3	2	3
U29	2	3	2	2	2	3	1	2
U30	2	3	2	3	2	3	2	3

Second experiment: avatar condition satisfaction statements

Avatar Condition Satisfaction Statements								
User	S1	S2	S3	S4	S5	S6	S7	S8
U1	4	3	3	3	3	3	3	3
U2	4	3	3	4	3	4	3	4
U3	4	4	4	3	3	3	2	3
U4	4	3	3	3	3	3	2	2
U5	3	3	3	4	4	4	3	4
U6	3	3	3	3	3	3	2	3
U7	3	3	3	3	3	3	2	3
U8	4	4	3	3	3	3	3	3
U9	3	3	3	3	3	4	3	3
U10	4	4	3	3	3	3	2	3
U11	4	3	3	4	3	4	3	3
U12	4	3	3	3	4	3	3	2
U13	4	4	3	3	3	4	3	3
U14	4	4	3	3	3	3	2	3
U15	4	4	4	3	3	3	3	3
U16	4	4	3	3	3	3	3	3
U17	3	3	3	3	3	4	3	3
U18	3	3	3	3	3	3	2	3
U19	3	3	3	4	4	4	3	4
U20	3	3	3	3	3	3	2	3
U21	4	4	3	3	3	4	3	3
U22	4	4	3	3	3	3	3	3
U23	4	4	3	4	3	3	2	3
U24	4	4	3	4	2	3	3	3
U25	3	3	3	3	3	3	2	3
U26	4	4	3	3	3	3	3	3
U27	4	4	3	4	4	4	3	3
U28	4	4	3	3	3	3	3	3
U29	4	4	3	4	2	3	3	3
U30	3	3	3	4	3	3	2	3

Second experiment: earcons condition satisfaction statements

Earcons Condition Satisfaction Statements								
User	S1	S2	S3	S4	S5	S6	S7	S8
U1	2	2	1	2	2	3	2	2
U2	2	3	1	1	1	3	3	2
U3	3	2	1	1	2	3	2	2
U4	2	3	1	2	1	3	1	2
U5	3	3	1	2	2	2	1	2
U6	3	3	1	2	1	3	2	1
U7	3	3	1	2	1	3	1	1
U8	2	2	2	1	2	3	2	3
U9	3	3	2	1	2	2	2	2
U10	3	2	2	2	1	3	1	2
U11	3	3	1	1	2	3	2	2
U12	2	2	1	2	2	3	1	2
U13	3	3	2	1	2	3	1	2
U14	3	3	2	2	2	3	2	2
U15	2	2	2	2	1	3	1	1
U16	3	2	1	2	1	3	1	1
U17	2	2	2	1	2	2	2	2
U18	3	2	2	2	2	2	1	1
U19	3	3	1	2	1	3	1	2
U20	3	3	1	2	2	3	2	2
U21	3	3	2	1	1	3	2	2
U22	2	3	1	2	1	3	2	1
U23	3	2	1	2	1	3	1	1
U24	3	3	2	2	1	3	2	3
U25	3	3	2	1	2	3	2	2
U26	3	3	1	2	1	3	2	2
U27	2	3	2	2	2	2	2	2
U28	3	3	2	1	1	3	2	2
U29	3	3	1	2	2	3	1	2
U30	3	3	2	1	1	3	2	2

Second experiment: classic music condition satisfaction statements

Classic Music Condition Satisfaction Satements								
User	S1	S2	S3	S4	S5	S6	S7	S8
U1	3	3	3	3	3	3	2	3
U2	3	3	3	3	3	3	3	3
U3	3	3	3	3	3	3	2	3
U4	3	3	3	3	3	2	2	3
U5	3	3	3	3	3	3	2	3
U6	3	3	3	2	3	3	2	3
U7	3	2	3	3	2	3	2	3
U8	3	3	3	3	3	3	2	3
U9	3	3	3	3	3	3	2	3
U10	3	3	3	3	3	3	2	3
U11	3	3	3	3	3	3	1	3
U12	3	3	3	3	3	3	2	3
U13	3	3	3	3	3	3	2	3
U14	3	3	3	3	3	3	2	3
U15	3	3	3	2	3	2	2	2
U16	3	3	3	3	3	3	1	3
U17	3	3	3	3	3	3	2	2
U18	3	2	3	3	3	3	2	3
U19	3	2	3	3	2	3	1	3
U20	3	3	3	3	3	3	2	3
U21	3	3	3	2	3	2	2	2
U22	3	2	3	3	2	3	2	3
U23	3	3	3	3	4	4	1	3
U24	3	2	3	3	2	3	2	3
U25	3	3	3	3	3	3	1	3
U26	3	3	3	3	3	3	2	3
U27	3	3	3	3	3	3	2	3
U28	3	2	3	3	3	3	2	3
U29	4	3	3	3	3	4	2	3
U30	3	3	3	3	3	3	2	3

Second experiment: speech condition satisfaction statements

Speech Condition Satisfaction Statements							
User	S1	S2	S3	S4	S5	S6	S7
U1	3	3	2	3	3	3	3
U2	4	3	3	3	3	3	2
U3	3	3	3	3	3	3	2
U4	3	3	3	3	3	3	1
U5	4	4	3	3	3	4	3
U6	3	3	3	3	3	3	1
U7	3	3	3	3	3	3	3
U8	3	3	3	3	3	3	3
U9	4	3	3	3	3	3	1
U10	3	3	3	3	3	3	2
U11	3	3	3	3	3	3	1
U12	3	2	3	3	3	2	1
U13	4	3	3	3	3	3	3
U14	3	3	3	3	3	3	2
U15	3	3	3	3	3	3	3
U16	3	3	3	3	3	3	2
U17	3	3	3	2	2	3	2
U18	3	3	3	3	2	2	1
U19	4	3	3	3	3	3	3
U20	3	3	3	3	2	3	2
U21	3	3	3	3	3	3	3
U22	4	3	3	3	3	3	2
U23	3	3	3	3	3	3	2
U24	3	3	3	3	3	3	2
U25	3	3	3	3	2	3	1
U26	3	2	3	3	3	3	3
U27	3	3	3	3	3	3	3
U28	3	3	3	3	3	3	2
U29	3	3	3	3	3	3	2
U30	3	2	3	3	3	3	2

Second experiment: nmmc condition satisfaction statements

NMMC Condition Satisfaction Statements								
User	S1	S2	S3	S4	S5	S6	S7	S8
U1	3	3	3	3	2	3	1	2
U2	4	3	2	2	1	2	2	2
U3	3	2	3	2	2	3	1	2
U4	2	2	3	2	2	2	2	2
U5	3	3	2	3	3	3	2	3
U6	3	2	3	2	2	3	2	2
U7	3	3	3	2	2	2	1	2
U8	3	3	3	2	2	3	1	2
U9	3	3	2	2	2	3	1	2
U10	3	3	3	2	2	3	1	2
U11	3	2	3	3	3	3	2	3
U12	4	3	2	2	2	2	1	2
U13	2	3	3	2	3	3	1	2
U14	2	3	3	2	3	3	1	3
U15	2	2	3	2	2	2	2	2
U16	3	3	2	2	2	3	2	2
U17	3	3	2	2	2	3	1	2
U18	3	3	3	2	2	3	1	2
U19	3	3	3	2	3	2	1	3
U20	4	3	3	2	3	3	1	3
U21	3	3	3	2	3	2	1	3
U22	4	2	3	2	1	3	1	2
U23	3	3	3	2	3	2	1	2
U24	3	2	2	3	2	3	1	2
U25	3	2	2	2	2	2	2	2
U26	3	3	3	3	2	3	2	3
U27	4	2	2	3	3	3	2	3
U28	3	3	2	2	2	2	2	2
U29	2	3	3	2	2	2	1	2
U30	3	2	2	2	3	3	1	3

Second experiment: user experience data

User Experience AttrakDiff instrument (Word-Pairs)							
Category	Pair of words	Auditory	AVATAR	Earcons	Speech	Music	NMMC
PQ	Technical - Human	0.40	1.77	-1.20	1.10	0.50	-0.77
	Complicated - Simple	0.57	2.10	-0.70	1.30	0.87	1.40
	Impractical - Practical	0.60	2.20	-0.56	1.00	0.93	1.47
	cumbersome- Straightforward	023	1.83	-1.03	1.23	0.33	1.33
	Unpredictable- Predictable	0.50	1.40	-0.40	1.03	0.17	1.10
	Confusing - Clear	0.50	1.87	-0.06	1.07	0.67	1.20
	Unruly - Manageable	0.73	1.90	-0.76	1.03	0.37	1.13
HQ-I	Isolating-Connective	0.57	1.43	-0.33	1.20	0.90	-1.30
	Unprofessional-Professional	0.53	2.23	-0.20	0.93	1.03	-0.67
	Tacky-Stylish	0.40	2.33	-0.63	1.07	1.07	-0.73
	Cheap-Premium	0.53	2.43	-0.46	1.17	0.90	-0.93
	Alienating-Integrating	0.23	2.20	-0.73	0.97	0.67	-0.40
	Separates me-Brings me closer	0.67	1.60	-0.83	0.40	0.20	-0.53
	Unpresentable-Presentable	0.60	1.80	-0.8	1.00	0.97	0.00
HQ-S	Conventional-Inventing	0.73	2.37	0.13	0.93	0.63	-2.23
	Unimaginative-Creative	0.37	1.80	0.10	0.93	0.67	-0.53
	Cautious-Bold	0.63	2.00	-0.13	0.53	0.67	-0.83
	Conservative-Innovative	0.30	2.30	0.10	0.93	1.10	-1.97
	Dull-Captivating	0.33	2.27	-0.40	0.90	0.90	-1.57
	Undemanding-Challenging	0.37	1.43	0.03	0.77	0.37	-1.10
	Ordinary-Novel	0.80	2.23	0.10	0.90	1.03	-1.63
ATT	Unpleasant-Pleasant	0.67	1.80	-0.80	1.13	1.03	-0.13
	Ugly-Attractive	0.80	2.37	-1.10	1.00	1.07	-157
	Disagreeable-Likable	0.53	2.13	-0.70	1.00	1.03	-0.60
	Rejecting-Inviting	0.53	2.13	-1.10	0.77	1.00	-23
	Bad - Good	0.37	2.13	-1.13	0.87	1.07	0.20
	Repelling - Appealing	0.47	1.80	-0.48	0.87	1.07	-0.43
	Discouraging - Motivation	0.40	2.37	-1.00	1.07	1.13	-1.57

Second experiment: user affective state, SAM (valence, arousal and dominance) data

	Second experimental phase (Affective state)																	
	Auditory			Avatar			Earcons			Music			Speech			NMMC		
User	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance	Valence	Arousal	Dominance
U1	4	4	4	6	6	7	2	8	2	7	5	5	6	5	5	5	5	5
U2	5	6	5	7	7	7	4	4	3	6	3	6	5	5	5	4	5	5
U3	4	5	6	8	7	6	1	9	1	6	3	5	6	6	5	4	4	4
U4	5	5	3	8	7	7	3	6	4	6	5	5	5	6	5	3	4	4
U5	4	6	5	7	6	6	3	6	3	6	6	6	6	5	5	4	5	3
U6	5	5	5	8	8	7	4	6	2	6	7	5	5	5	5	5	4	4
U7	4	5	4	6	5	5	2	8	3	5	5	5	5	5	4	3	4	4
U8	4	6	4	6	5	5	2	7	1	6	5	5	6	6	5	6	4	5
U9	4	4	2	8	8	8	3	5	3	6	4	4	5	5	5	6	5	4
U10	4	6	5	7	6	5	4	5	4	5	3	4	6	5	5	5	4	4
U11	5	5	5	7	7	6	5	5	2	7	7	7	5	6	5	4	5	4
U12	4	5	5	7	6	6	4	4	4	7	7	8	5	5	5	5	4	4
U13	5	4	3	6	6	5	4	4	4	6	5	5	5	5	5	5	3	4
U14	4	3	3	8	7	7	2	8	2	4	6	3	6	5	5	5	4	3
U15	5	4	4	7	6	5	3	6	2	6	4	5	5	5	4	5	3	4
U16	4	5	5	7	6	5	4	6	3	7	7	6	6	5	4	4	4	4
U17	4	5	4	6	5	5	3	5	4	7	6	7	5	5	4	5	3	3
U18	5	5	5	6	6	5	4	4	4	5	5	4	7	5	5	4	4	3
U19	4	4	5	7	6	6	2	6	1	5	4	4	6	6	4	4	6	2
U20	4	4	4	7	5	5	4	4	3	6	5	5	5	5	5	4	4	4
U21	5	5	5	6	6	5	4	6	3	7	5	5	5	5	5	5	5	4
U22	4	6	5	8	7	7	2	6	2	6	6	6	6	5	6	5	5	4
U23	5	5	1	6	6	5	4	5	1	7	8	5	6	5	5	4	3	3
U24	4	6	6	9	8	8	4	5	4	6	5	6	6	5	5	5	4	4
U25	5	5	5	7	6	6	4	6	4	6	5	5	5	5	5	5	5	5
U26	4	6	5	7	6	6	1	8	1	7	6	6	7	5	5	5	6	2
U27	4	5	5	7	6	6	2	6	3	5	6	5	6	6	4	4	3	4
U28	5	4	4	9	8	8	4	5	5	9	8	8	5	5	5	3	3	2
U29	5	5	5	6	6	5	2	3	1	8	7	7	5	5	5	5	4	4
U30	4	6	2	8	7	7	4	6	2	6	4	5	6	6	4	4	3	3

Second experiment: user affective state biofeedback (skin conductance) data

EDA (µS)						
User	Auditory	Avatar	Earcons	Music	Speech	NMMC
U1	3.900	4.937	4.498	3.577	4.192	2.243
U2	1.745	2.917	2.233	2.220	2.220	1.402
U3	3.165	5.690	5.902	2.343	1.875	1.838
U4	15.940	15.877	15.948	16.204	16.228	14.505
U5	5.125	5.101	5.718	4.680	4.682	3.642
U6	5.939	6.113	5.830	5.557	5.825	3.710
U7	12.191	13.637	12.980	11.876	12.470	10.722
U8	12.366	12.501	13.606	10.817	12.014	9.615
U9	6.870	7.605	7.587	6.306	6.594	4.732
U10	14.756	14.742	15.421	13.106	10.106	9.502
U11	11.850	12.918	12.910	11.773	9.899	8.337
U12	3.917	5.790	4.671	2.508	3.068	1.680
U13	10.744	12.700	11.764	11.024	13.044	10.108
U14	4.803	5.093	5.222	4.676	4.714	3.843
U15	12.534	12.905	12.641	11.030	12.300	10.366
U16	0.750	0.801	0.727	0.658	0.713	0.418
U17	0.574	0.704	0.751	0.587	0.564	0.256
U18	0.376	0.676	0.379	0.355	0.340	0.091
U19	7.614	9.750	9.433	6.781	8.046	7.554
U20	0.646	0.763	0.726	0.564	0.601	0.370
U21	10.257	11.584	13.715	13.142	13.040	10.738
U22	6.954	9.378	8.346	6.463	7.275	5.616
U23	12.436	13.212	14.782	14.353	12.242	11.748
U24	8.156	8.580	8.382	8.802	7.184	6.267
U25	14.644	14.298	16.461	15.930	15.817	14.172
U26	3.602	3.727	3.801	3.658	3.713	3.558
U27	11.190	12.625	12.964	11.875	12.470	10.720
U28	10.533	10.507	12.631	12.029	12.300	10.383
U29	4.225	5.376	5.937	3.577	4.287	2.700
U30	15.648	14.329	16.462	15.936	15.826	16.177

Second experiment: user affective state, biofeedback (skin temperature) data

7.4.1.1 ***Skin temperature ºC***						
User	**Auditory**	**Avatar**	**Earcons**	**Music**	**Speech**	**NMMC**
U1	34.1375	33.8336	34.2217	34.1933	34.1925	34.1366
U2	35.6654	35.5450	36.0105	35.7217	35.6950	35.7815
U3	35.2144	35.0844	35.3431	35.2693	35.1911	35.2690
U4	36.4132	36.1220	36.4145	36.4001	36.3147	36.3713
U5	36.2790	36.5489	36.6053	36.3378	36.0285	36.4623
U6	35.0840	35.3512	35.4761	35.2336	34.9391	35.2138
U7	33.5446	33.7105	34.3431	33.9044	34.1469	34.4707
U8	35.4120	35.0486	35.6247	35.4213	35.9549	35.5683
U9	34.5826	34.2656	34.5487	34.5010	34.3690	34.3096
U10	35.0861	35.5022	35.7822	35.2453	35.4536	35.6254
U11	34.0141	34.0582	34.7245	33.8746	34.2493	34.3478
U12	34.8253	35.3970	36.0627	34.5216	34.6522	34.8299
U13	35.9065	34.8992	35.3482	34.9754	34.8597	34.7615
U14	36.2756	36.5068	36.8423	36.5505	36.7682	36.7087
U15	32.9788	33.2233	33.3638	33.4701	33.3710	33.2070
U16	33.7596	34.5267	34.4754	33.2007	33.8753	33.4757
U17	32.0735	32.6219	33.3953	32.0796	32.3928	32.4784
U18	33.2378	32.9587	33.8347	33.3558	33.1578	33.4608
U19	34.9128	34.3172	34.5913	34.4571	34.8216	34.7801
U20	35.7931	35.2962	35.9683	35.4214	35.6148	35.2329
U21	35.7931	35.7145	35.9683	35.4214	35.6148	35.2329
U22	36.1595	36.7361	36.0673	35.7746	36.1718	35.0693
U23	31.4774	31.2578	31.5203	31.4250	31.3028	31.3863
U24	31.3869	31.0700	31.3376	31.4160	31.2098	31.2542
U25	35.3175	35.1335	35.3775	35.2171	35.3587	35.3988
U26	35.2285	35.6074	35.5913	34.4571	35.2676	32.5340
U27	33.8446	33.7117	34.0318	33.9044	33.8726	33.7286
U28	33.3788	33.2233	33.3973	33.4701	33.3591	33.3700
U29	34.4689	34.1803	34.5440	33.8966	34.0933	34.4052
U30	35.3175	35.1335	35.3775	35.2171	35.2374	35.2900

REFERENCES

[1] Alseid, Marwan, and Dimitrios Rigas. "Efficiency of multimodal metaphors in the presentation of learning information." In BCS HCI (2), pp. 107-110. 2008.

[2] Rigas, Dimitris, and Marwan Alseid. "Multi-modal aided Presentation of Learning Information: A Usability Comparative Study." In Proceedings of IADIS International Conference on Interfaces and Human-Computer Interaction, pp. 234-238. 2008.

[3] Rigas, Dimitrios, and Khaled Ayad. "Using edutainment in e-learning application: an empirical study." International Journal of Computers 4, no. 1 (2010): 36-43.

[4] RIGAS, DIMITRIOS, and MOHAMED SALLAM. "An Empirical Study on E-learning Note-Taking Platform Using Multimodal Metaphors." Technology 18 (2008): 19.

[5] Hertzum, Morten, and Torkil Clemmensen. "How do usability professionals construe usability?." International Journal of Human-Computer Studies 70, no. 1 (2012): 26-42.

[6] DILLON, Andrew. "Beyond usability: Process, outcome, and affect in human computer interactions." Canadian journal of information and library science 26, no. 4 (2001): 57-69.

[7] Battarbee, Katja, and Ilpo Koskinen. "Co-experience: user experience as interaction." CoDesign 1, no. 1 (2005): 5-18.

[8] Bevan, Nigel. Extending quality in use to provide a framework for usability measurement. Springer Berlin Heidelberg, 2009.

[9] Norman, Donald A. The design of future things. Basic Books, 2007.

[10] Bevan, Nigel. "International standards for usability should be more widely used." Journal of Usability Studies 4, no. 3 (2009): 106-113.

[11] Petrie, Helen, and Nigel Bevan. "The evaluation of accessibility, usability and user experience." The universal access handbook (2009): 10-20.

[12] Bevan, Nigel. Extending quality in use to provide a framework for usability measurement. Springer Berlin Heidelberg, 2009.

[13] Hassenzahl, Marc. "The thing and I: understanding the relationship between user and product." In Funology, pp. 31-42. Springer Netherlands, 2005.

[14] Law, Effie Lai-Chong, Virpi Roto, Marc Hassenzahl, Arnold POS Vermeeren, and Joke Kort. "Understanding, scoping and defining user experience: a survey approach." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 719-728. ACM, 2009.

[15] Hassenzahl, Marc, Michael Burmester, and Franz Koller. "AttrakDiff™." Information available at http://www. attrakdiff. de (Accessed: 06-Jun-2011).

[16] Bradley, Margaret M., and Peter J. Lang. "Measuring emotion: the self-assessment manikin and the semantic differential." Journal of behaviour therapy and experimental psychiatry 25, no. 1 (1994): 49-59.

[17] Affectiva, “Understanding EDA Data”, Internet Resource (Accessed: 06-Jun-2011) http: //www. affectiva.com/q-sensor .

[18] Downes, Stephen. "Feature: E-learning 2.0." Elearn magazine 2005, no. 10 (2005): 1.

[19] Zhang, Dongsong, and Jay F. Nunamaker. "Powering e-learning in the new millennium: an overview of e-learning and enabling technology." Information Systems Frontiers 5, no. 2 (2003): 207-218.

[20] Sun, Pei-Chen, Ray J. Tsai, Glenn Finger, Yueh-Yang Chen, and Dowming Yeh. "What drives a successful e-Learning? An empirical investigation of the critical factors influencing learner satisfaction." Computers & Education 50, no. 4 (2008): 1183-1202.

[21] Archibald, Thomas N., and Jonathan Poltrack. "ADL: Preliminary systems and content integration research within the next generation learning environment." In The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC), vol. 2011, no. 1. National Training Systems Association, 2011.

[22] Garrison, D. Randy. E-learning in the 21st century: A framework for research and practice. Taylor & Francis, 2011.

[23] Wisher, Robert A., and J. Dexter Fletcher. "The case for advanced distributed learning." Information and Security 14 (2004): 17-25.

[24] Moore, Joi L., Camille Dickson-Deane, and Krista Galyen. "e-Learning, online learning, and distance learning environments: Are they the same?." The Internet and Higher Education 14, no. 2 (2011): 129-135.

[25] Cohen, Eli, and Malgorzata Nycz. "Learning objects and e-learning: An informing science perspective." Interdisciplinary Journal of E-Learning and Learning Objects 2, no. 1 (2006): 23-34.

[26] Wiley, D. A. "Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In DA Wiley (Ed.), The Instructional Use of Learning Objects." pp. http://reusability. org/read/chapters/wiley. doc (Accessed: 12-Feb-2012) .

[27] Moreno, Roxana, and Richard Mayer. "Interactive multimodal learning environments." Educational Psychology Review 19, no. 3 (2007): 309-326.

[28] Fletcher, J. D. "Does this stuff work? A review of technology used to teach." (2003): 10-14.

[29] Hall, Brandon. The Web-Based Training Cookbook with Cdrom. John Wiley & Sons, Inc., 1997.

[30] Kruse, Kevin. "The benefits and drawbacks of e-learning." available at http://www.e-learningguru. com (Accessed: March 5, 2011).

[31] Hobbs, M., E. Brown, and M. Gordon. "Using a virtual world for transferable skills in gaming education." Innovation in teaching and learning in information and computer sciences 5, no. 3 (2006).

[32] SHUDAYFAT, Eman Ahmad, Florica Moldoveanu, Alin Moldoveanu, and Alexandru Grădinaru."VIRTUAL REALITY-BASED BIOLOGY LEARNING MODULE." In Conference proceedings of ." eLearning and Software for Education"(eLSE), no. 02, pp. 621-626. 2013.

[33] Pringle, Jamie K. "Educational environmental geoscience e-gaming to provide stimulating and effective learning." Planet 27, no. 1 (2013): 21-28.

[34] Paechter, Manuela, Brigitte Maier, and Daniel Macher. "Students' expectations of, and experiences in e-learning: Their relation to learning achievements and course satisfaction." Computers & Education 54, no. 1 (2010): 222-229.

[35] Vrasidas, Charalambos. "Issues of pedagogy and design in e-learning systems." In Proceedings of the 2004 ACM symposium on Applied computing, pp. 911-915. ACM, 2004.

[36] Thorndike, E. L., B. F. Skinner, and Albert Bandura. "Learning Theories and Pedagogy." (2010).

[37] Nichols, Mark. "A theory for E-Learning." Educational technology & society 6, no. 2 (2003): 1-10.

[38] Holmes, Geraldine, and Michele Abington-Cooper. "Pedagogy vs. andragogy: A false dichotomy?." Pedagogy 26, no. 2 (2000).

[39] Mehanna, Wassila Naamani. "e-Pedagogy: the pedagogies of e-learning." Research in Learning Technology 12, no. 3 (2004).

[40] Kuriloff, Peshe. "Breaking the Barriers of Time and Space More Effective Teaching Using e-Pedagogy." Innovate Journal of Online Education 2, no. 1 (2005): 64-72.

[41] Dictionary, Oxford English. "Oxford English Dictionary." Oxford University Press (2013).

[42] Jaimes, Alejandro, and Nicu Sebe. "Multimodal human–computer interaction: A survey." Computer vision and image understanding 108, no. 1 (2007): 116-134.

[43] Brewster, Stephen A. "Non-speech auditory output." Human Computer Interaction Handbook, Lawrence Erlbaum Associates, Mahwah, NJ (2002): 220-239.

[44] Jadin, Tanja, Astrid Gruber, and Bernad Batinic. "Learning with E-lectures: The Meaning of Learning Strategies." Educational Technology & Society 12, no. 3 (2009): 282-288.

[45] Oviatt, Sharon, Rachel Coulston, and Rebecca Lunsford. "When do we interact multimodally?: cognitive load and multimodal communication patterns." In Proceedings of the 6th international conference on Multimodal interfaces, pp. 129-136. ACM, 2004.

[46] Rachovides, Dorothy, Zoë Swiderski, and Alan P. Parkes. "Metaphors and Multimodal Interaction." R GN E P CIES (2001): 125.

[47] K. JOKINEN and A. RAIKE, Multimodality–technology, visions and demands for the future, Citeseer, 2003, pp. 239–251.

[48] Oviatt, Sharon. "Multimodal interfaces." The human-computer interaction handbook: Fundamentals, evolving technologies and emerging applications (2003): 286-304.

[49] Mayer, Richard E. "Cognitive theory of multimedia learning." The Cambridge handbook of multimedia learning (2005): 31-48.

[50] Blattner, Meera M., Denise A. Sumikawa, and Robert M. Greenberg. "Earcons and icons: Their structure and common design principles." Human–Computer Interaction 4, no. 1 (1989): 11-44.

[51] Walker, Bruce N., and Gregory Kramer. "Ecological psychoacoustics and auditory displays: Hearing, grouping, and meaning making." Ecological psychoacoustics (2004): 150-175.

[52] Petocz, Agnes, Peter E. Keller, and Catherine J. Stevens. "Auditory warnings, signal-referent relations, and natural indicators: re-thinking theory and application." Journal of experimental psychology: Applied 14, no. 2 (2008): 165.

[53] Walker, Bruce N., and Jeffrey Lindsay. "Navigation performance with a virtual auditory display: Effects of beacon sound, capture radius, and practice." Human Factors: The Journal of the Human Factors and Ergonomics Society 48, no. 2 (2006): 265-278.

[54] Walker, Bruce N., Jeffrey Lindsay, Amanda Nance, Yoko Nakano, Dianne K. Palladino, Tilman Dingler, and Myounghoon Jeon. "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus." Human Factors: The Journal of the Human Factors and Ergonomics Society 55, no. 1 (2013): 157-182.

[55] Brewster, Stephen A., Peter C. Wright, and Alistair DN Edwards. "An evaluation of earcons for use in auditory human-computer interfaces." In Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems, pp. 222-227. ACM, 1993.

[56] Brewster, Stephen A., Peter C. Wright, and Alistair DN Edwards. "A detailed investigation into the effectiveness of earcons." In In Proceedings of ICAD'92 (Santa Fe Institute, Santa Fe) Addison-Wesley. (1994): 471-498.

[57] Cole, Richard C., and Ed Schwartz. “Virginia Tech multimedia music dictionary.” Virginia Tech, 2012.

[58] Wilson, James W. The musical cyclopedia: being a collection of the most approved English, Scottish, and Irish songs. Parry & Co., 1852.

[59] Patterson, R. D., and T. F. Mayfield. "Auditory warning sounds in the work environment [and Discussion]." Philosophical Transactions of the Royal Society of London. B, Biological Sciences 327, no. 1241 (1990): 485-492.

[60] Duarte, Tiago, Rafael Prikladnicki, Fabio Calefato, and Filippo Lanubile. "Speech Recognition for Voice-Based Machine Translation." IEEE software 31, no. 1 (2014): 26-31.

[61] Lewis, Michael, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, eds. Handbook of emotions. Guilford Press, 2010.

[62] Stevenson, Angus, and Maurice Waite, eds. Concise Oxford English Dictionary: Book & CD-ROM Set. Oxford University Press, 2011.

[63] Kim, Jonghwa, and Elisabeth André. "Emotion recognition based on physiological changes in music listening." Pattern Analysis and Machine Intelligence, IEEE Transactions on 30, no. 12 (2008): 2067-2083.

[64] Jerritta, S., M. Murugappan, R. Nagarajan, and Khairunizam Wan. "Physiological signals based human emotion recognition: a review." In Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on, pp. 410-415. IEEE, 2011.

[65] Chang, Chuan-Yu, Jun-Ying Zheng, and Chi-Jane Wang. "Based on support vector regression for emotion recognition using physiological signals." In Neural Networks (IJCNN), The 2010 International Joint Conference on, pp. 1-7. IEEE, 2010.

[66] Panksepp, Jaak, and Günther Bernatzky. "Emotional sounds and the brain: the neuro-affective foundations of musical appreciation." Behavioural processes 60, no. 2 (2002): 133-155.

[67] Savan, Anne. "The effect of background music on learning." Psychology of Music 27, no. 2 (1999): 138-146.

[68] Kämpfe, Juliane, Peter Sedlmeier, and Frank Renkewitz. "The impact of background music on adult listeners: A meta-analysis." Psychology of Music 39, no. 4 (2011): 424-448.

[69] De Groot, Annette. "Effects of stimulus characteristics and background music on foreign language vocabulary learning and forgetting." Language Learning 56, no. 3 (2006): 463-506.

[70] Oxford Online Dictionary. “Avatar.” available at http://www. Oxforddictionaries. com/ definition/ english/ avatar (Accessed: 23-Jun-2011).

[71] Burrows, Roger, and Mike Featherstone, eds.“Cyberspace/Cyberbodies/Cyberpunk: Cultures of Technological Embodiment.” Sage, 2000.

[72] Gerhard, Michael, David Moore, and Dave Hobbs. "Embodiment and copresence in collaborative interfaces." International Journal of Human-Computer Studies 61, no. 4 (2004): 453-480.

[73] Borzo, Jeanette. "Almost human." The Wall Street Journal 24 (2004).

[74] Sheth, Raj. "Avatar technology: giving a face to the e-learning interface." The eLearning Developers’ Journal (2003): 1-10.

[75] Alseid, Marwan, and Dimitrios Rigas. "Utilising Multimodal interaction metaphors in e-learning applications: an experimental study." In Advanced Information Networking and Applications, 2009. AINA'09. International Conference on, pp. 945-950. IEEE, 2009.

[76] Blascovich, Jim. "A theoretical model of social influence for increasing the utility of collaborative virtual environments." In Proceedings of the 4th international Conference on Collaborative Virtual Environments, pp. 25-30. ACM, 2002.

[77] Benford, Steve, John Bowers, Lennart E. Fahlén, Chris Greenhalgh, and Dave Snowdon. "User embodiment in collaborative virtual environments." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 242-249. ACM Press/Addison-Wesley Publishing Co., 1995.

[78] Gerhard, Michael. "A hybrid avatar/agent model for educational CVEs." PhD diss., PhD Thesis, Leeds Metropolitan University, 2003.

[79] Maher, Mary Lou. "Designers and collaborative virtual environments." In Collaborative design in virtual environments, pp. 3-15. Springer Netherlands, 2011.

[80] Oestreicher, Klaus, Joanne Kuzma, and Dorothy Yen. "Avatar Supported Learning in a Virtual University." Worcester Journal of Learning and Teaching 4 (2010): 1-8.

[81] Costabile, Maria Francesca. "Usability in the software life cycle." Handbook of software engineering and knowledge engineering 1 (2001): 179-192.

[82] Earthy, Jonathan, Brian Sherwood Jones, and Nigel Bevan. "The improvement of human-centred processes—facing the challenge and reaping the benefit of ISO 13407." International Journal of Human-Computer Studies 55, no. 4 (2001): 553-585.

[83] Zaharias, Panagiotis, and Angeliki Poylymenakou. "Developing a usability evaluation method for e-learning applications: Beyond functional usability." Intl. Journal of Human–Computer Interaction 25, no. 1 (2009): 75-98.

[84] Nielsen, Jakob. "The usability engineering life cycle." Computer 25, no. 3 (1992): 12-22.

[85] Ardito, Carmelo, Maria Francesca Costabile, Marilena De Marsico, Rosa Lanzilotti, Stefano Levialdi, Teresa Roselli, and Veronica Rossano. "An approach to usability evaluation of e-learning applications." Universal access in the information society 4, no. 3 (2006): 270-283.

[86] Bevan, Nigel. "UX, usability and ISO standards." In 26 th Annual Conference on Computer Human Interaction (CHI), ACM SIGCHI. 2008.

[87] Feldstein, Michael, and Lisa Neal. "Designing usable, self-paced e-learning courses: a practical guide." eLearn 2006, no. 8 (2006): 1.

[88] Lee, Sangwon, and Richard J. Koubek. "Users' perceptions of usability and aesthetics as criteria of pre–and post–use preferences." European Journal of Industrial Engineering 6, no. 1 (2012): 87-117.

[89] Ferré, Xavier, Natalia Juristo, Helmut Windl, and Larry Constantine. "Usability basics for software developers." IEEE software 18, no. 1 (2001): 22-29.

[90] Sethi, Monika, and Anju Sharma. "Information System and System Development Life Cycle." Software Development Techniques for Constructive Information Systems Design (2013): 118.

[91] Costabile, Maria Francesca, Marilena De Marsico, Rosa Lanzilotti, Vito Leonardo Plantamura, and Teresa Roselli. "On the usability evaluation of e-learning applications." In System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on, pp. 6b-6b. IEEE, 2005.

[92] Harms, Michael, and Jeremy Adams. "Usability and design considerations for computer-based learning and assessment." In Meeting of the American Educational Research Association (AERA), March. 2008.

[93] Norman, Donald A. “Emotional design: Why we love (or hate) everyday things.” Basic books, 2004.

[94] Wechsung, Ina, Anja B. Naumann, and Robert Schleicher. "Views on Usability and User Experience: from Theory and Practice." Deutsche Telekom Laboratories (2008).

[95] Vermeeren, Arnold POS, Effie Lai-Chong Law, Virpi Roto, Marianna Obrist, Jettie Hoonhout, and Kaisa Väänänen-Vainio-Mattila. "User experience evaluation methods: current state and development needs." In Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries, pp. 521-530. ACM, 2010.

[96] Kaye, Joseph'Jofish. "Evaluating experience-focused HCI." In CHI'07 extended abstracts on Human factors in computing systems, pp. 1661-1664. ACM, 2007.

[97] Law, Effie Lai-Chong, Virpi Roto, Marc Hassenzahl, Arnold POS Vermeeren, and Joke Kort. "Understanding, scoping and defining user experience: a survey approach." In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 719-728. ACM, 2009.

[98] Forlizzi, Jodi, and Katja Battarbee. "Understanding experience in interactive systems." In Proceedings of the 5th conference on Designing interactive systems: processes, practices, methods, and techniques, pp. 261-268. ACM, 2004.

[99] Hassenzahl, Marc, and Noam Tractinsky. "User experience-a research agenda." Behaviour & Information Technology 25, no. 2 (2006): 91-97.

[100] Russell, James A. "Core affect and the psychological construction of emotion." Psychological review 110, no. 1 (2003): 145.

[101] Mehrabian, Albert, and James A. Russell. "A verbal measure of information rate for studies in environmental psychology." Environment and Behavior (1974).

[102] Gunes, Hatice, and Maja Pantic. "Automatic, dimensional and continuous emotion recognition." International Journal of Synthetic Emotions (IJSE) 1, no. 1 (2010): 68-99.

[103] Broekens, Joost. "In defense of dominance: PAD usage in computational representations of affect." International Journal of Synthetic Emotions (IJSE) 3, no. 1 (2012): 33-42.

[104] Schutz, Paul A., and Reinhard Pekrun. "Introduction to emotion in education."Emotion in education (2007): 3-10.

[105] Desmet, Pieter MA, and P. Hekkert. "The basis of product emotions." Pleasure with products, beyond usability (2002): 60-68.

[106] D'Hondt, Fabien, Maryse Lassonde, Olivier Collignon, Anne-Sophie Dubarry, Manon Robert, Simon Rigoulot, Jacques Honoré, Franco Lepore, and Henrique Sequeira. "Early brain-body impact of emotional arousal." Frontiers in human neuroscience 4 (2010).

[107] Sharma, Tanu, and Bhanu Kapoor. "Emotion estimation of physiological signals by using low power embedded system." In Proceedings of the Conference on Advances in Communication and Control Systems-2013. Atlantis Press, 2013.

[108] Cacioppo, John T., Gary G. Berntson, Jeff T. Larsen, Kirsten M. Poehlmann, and Tiffany A. Ito. "The psychophysiology of emotion." Handbook of emotions 2 (2000): 173-191.

[109] Sánchez-Navarro, Juan Pedro, José María Martínez-Selva, Ginesa Torrente, and Francisco Román. "Psychophysiological, behavioral, and cognitive indices of the emotional response: A factor-analytic study." The Spanish journal of psychology 11, no. 01 (2008): 16-25.

[110] Wikipedia. "Emotion."available at http: //en. wikipedia. org/ wiki/ Emotion (Accessed: 12-Jan-2011).

[111] Oxford Online Dictionary. "Emotion." available at http:// www. oxforddictionaries. com/ definition/ english/ emotion (Accessed: 15-Jan-2011).

[112] Solomon, Robert C. "The philosophy of emotions." Handbook of emotions (1993): 3-15. [113] STEARNS, PETER N. "History of Emotions." EMOTIONS: 17.

[114] Titchener, E. B. "A historical note on the James-Lange theory of emotion." The American Journal of Psychology (1914): 427-447.

[115] Damasio, Antonio, and Gil B. Carvalho. "The nature of feelings: evolutionary and neurobiological origins." Nature Reviews Neuroscience 14, no. 2 (2013): 143-152.

[116] Schachter, Stanley. "The interaction of cognitive and physiological determinants of emotional state." Advances in experimental social psychology 1 (1964): 49-80.

[117] Reisenzein, Rainer. "The Schachter theory of emotion: two decades later." Psychological bulletin 94, no. 2 (1983): 239.

[118] Ekman, Paul, Wallace V. Friesen, Maureen O'Sullivan, Anthony Chan, Irene Diacoyanni-Tarlatzis, Karl Heider, Rainer Krause et al. "Universals and cultural differences in the judgments of facial expressions of emotion." Journal of personality and social psychology 53, no. 4 (1987): 712.

[119] Lang, Peter J. "The emotion probe: Studies of motivation and attention." American psychologist 50, no. 5 (1995): 372.

[120] Sabini, John, and Maury Silver. "Ekman's basic emotions: Why not love and jealousy?." Cognition & Emotion 19, no. 5 (2005): 693-712.

[121] Ekman, Paul. "An argument for basic emotions." Cognition & Emotion 6, no. 3-4 (1992): 169-200.

[122] Ekman, Paul, and Daniel Cordaro. "What is meant by calling emotions basic." Emotion Review 3, no. 4 (2011): 364-370.

[123] Bradley, Margaret M., and Peter J. Lang. "Affective reactions to acoustic stimuli." Psychophysiology 37, no. 02 (2000): 204-215.

[124] Peter, Christian, and Antje Herbon. "Emotion representation and physiology assignments in digital systems." Interacting with Computers 18, no. 2 (2006): 139-170.

[125] Dalgleish, Tim, ed. Handbook of cognition and emotion. Chichester, UK: Wiley, 1999.

[126] Bradley, Margaret M., and Peter J. Lang. "Measuring emotion: Behavior, feeling, and physiology." Cognitive neuroscience of emotion 25 (2000): 49-59.

[127] Posner, Jonathan, James A. Russell, and Bradley S. Peterson. "The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology." Development and psychopathology 17, no. 03 (2005): 715-734.

[128] Ping, Heng Yu, Lili Nurliyana Abdullah, Alfian Abdul Halin, and Puteri Suhaiza Sulaiman. "A Study of Physiological Signals-based Emotion Recognition Systems." International Journal of Computers & Technology 11, no. 1 (2013): 2189-2196.

[129] Dirican, Ahmet Cengizhan, and Mehmet Göktürk. "Psychophysiological measures of human cognitive states applied in human computer interaction." Procedia Computer Science 3 (2011): 1361-1367.

[130] Kreibig, Sylvia D. "Autonomic nervous system activity in emotion: A review." Biological psychology 84, no. 3 (2010): 394-421.

[131] Mendes, Wendy Berry. "Assessing autonomic nervous system activity." Methods in social neuroscience (2009): 118-147.

[132] Salimpoor, Valorie N., Mitchel Benovoy, Gregory Longo, Jeremy R. Cooperstock, and Robert J. Zatorre. "The rewarding aspects of music listening are related to degree of emotional arousal." PloS one 4, no. 10 (2009): e7487.

[133] Lee, Nick, Amanda J. Broderick, and Laura Chamberlain. "What is 'neuromarketing'? A discussion and agenda for future research." International Journal of Psychophysiology 63, no. 2 (2007): 199-204.

[134] Dawson, Michael E., Anne M. Schell, and Diane L. Filion. "7 The Electrodermal System." Handbook of psychophysiology (2007): 159.

[135] Vitouch, Peter. "Psychophysiological methods in media research." In New horizons in media psychology, pp. 116-124. VS Verlag für Sozialwissenschaften, 1997.

[136] Winkielman, Piotr, and Kent C. Berridge. "Unconscious emotion." Current directions in psychological science 13, no. 3 (2004): 120-123.

[137] Berridge, Kent, and Piotr Winkielman. "What is an unconscious emotion?(The case for unconscious" liking")." Cognition & Emotion 17, no. 2 (2003): 181-211.

[138] Chalmers, Patricia A. "The role of cognitive theory in human–computer interface." Computers in Human Behavior 19, no. 5 (2003): 593-607.

[139] Rozell, E. J., and W. L. Gardner III. "Cognitive, motivation, and affective processes associated with computer-related performance: a path analysis."Computers in Human Behavior 16, no. 2 (2000): 199-222.

[140] Cannon, Walter B. "The interrelations of emotions as suggested by recent physiological researches." The American Journal of Psychology (1914): 256-282.

[141] Tähti, Marika, and Marketta Niemelä. "3e–expressing emotions and experiences." In Proc. WP9 Workshop on Innovative Approaches for Evaluating Affective Systems. HUMAINE (Human-Machine Interaction Network on Emotion). 2006.

[142] Russell, James A., Anna Weiss, and Gerald A. Mendelsohn. "Affect grid: a single-item scale of pleasure and arousal." Journal of personality and social psychology 57, no. 3 (1989): 493.

[143] Desmet, Pieter, Kees Overbeeke, and Stefan Tax. "Designing Products with Added Emotional Value: Development and Appllcation of an Approach for Research Through Design." The design journal 4, no. 1 (2001): 32-47.

[144] Isbister, Katherine, Kia Höök, Jarmo Laaksolahti, and Michael Sharp. "The sensual evaluation instrument: Developing a trans-cultural self-report measure of affect." International journal of human-computer studies 65, no. 4 (2007): 315-328.

[145] Boehner, Kirsten, Rogério DePaula, Paul Dourish, and Phoebe Sengers. "How emotion is made and measured." International Journal of Human-Computer Studies 65, no. 4 (2007): 275-291.

[146] Crawford, John R., and Julie D. Henry. "The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample." British Journal of Clinical Psychology 43, no. 3 (2004): 245-265.

[147] Watson, David, Lee A. Clark, and Auke Tellegen. "Development and validation of brief measures of positive and negative affect: the PANAS scales." Journal of personality and social psychology 54, no. 6 (1988): 1063.

[148] Bänziger, Tanja, Véronique Tran, and Klaus R. Scherer. "The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions." Poster presented at ISRE (2005). [149] T. BÄNZIGER and K. R. SCHERER, Introducing the geneva multimodal emotion portrayal (gemep) corpus, Blueprint for affective computing: A sourcebook (2010), pp. 271-294.

[150] Karapanos, Evangelos, Jean-Bernard Martens, and Marc Hassenzahl. "Reconstructing experiences through sketching." arXiv preprint arXiv:0912.5343 (2009).

[151] Desmet, Pieter. "Measuring emotion: Development and application of an instrument to measure emotional responses to products." In Funology, pp. 111-123. Springer Netherlands, 2005.

[152] Desmet, Pieter MA, Rick Porcelijn, and M. B. Van Dijk. "Emotional Design; Application of a research-based design approach." Knowledge, Technology & Policy 20, no. 3 (2007): 141-155.

[153] Sieverink, Linda. "Symbolic product meaning and commitments as predictors of product emotions." (2011).

[154] Hove, Siw Elisabeth, and Bente Anda. "Experiences from conducting semi-structured interviews in empirical software engineering research." In Software Metrics, 2005. 11th IEEE International Symposium, pp. 10-pp. IEEE, 2005.

[155] DiCicco-Bloom, Barbara, and Benjamin F. Crabtree. "The qualitative research interview." Medical education 40, no. 4 (2006): 314-321.

[156] Hole, Linda, and Oliver M. Williams. "The emotion sampling device (ESD)." InProceedings of the 21st British HCI Group Annual Conference on People and Computers: HCI but not as we know it-Volume 2, pp. 177-178. British Computer Society, 2007.

[157] Hassenzahl, Marc, Michael Burmester, and Franz Koller. "AttrakDiff™." Information available at http://www. attrakdiff. de / index-en.html (Accessed: 16-Feb-2012).

[158] Terzis, Vasileios, Christos N. Moridis, and Anastasios A. Economides. "Measuring instant emotions based on facial expressions during computer-based assessment." Personal and ubiquitous computing 17, no. 1 (2013): 43-52.

[159] NOLDUS, ”FaceReader™.” available at http://www.noldus.com/human-behavior-research/products/facereader (Accessed: 11-Feb-2012).

[160] Kivikangas, J. Matias, Guillaume Chanel, Ben Cowley, Inger Ekman, Mikko Salminen, Simo Järvelä, and Niklas Ravaja. "A review of the use of psychophysiological methods in game research." Journal of Gaming & Virtual Worlds 3, no. 3 (2011): 181-199.

[161] Jerritta, S., M. Murugappan, R. Nagarajan, and Khairunizam Wan. "Physiological signals based human emotion recognition: a review." In Signal Processing and its Applications (CSPA), 2011 IEEE 7th International Colloquium on, pp. 410-415. IEEE, 2011.

[162] Boucsein, Wolfram. Electrodermal activity. Springer, 2012.

[163] Park, Byungho. "Psychophysiology as a Tool for hci Research: Promises and Pitfalls." In Human-Computer Interaction. New Trends, pp. 141-148. Springer Berlin Heidelberg, 2009.

[164] Kim, Kyung Hwan, S. W. Bang, and S. R. Kim. "Emotion recognition system using short-term monitoring of physiological signals." Medical and biological engineering and computing 42, no. 3 (2004): 419-427.

[165] Kulic, Dana, and Elizabeth A. Croft. "Affective state estimation for human–robot interaction." Robotics, IEEE Transactions on 23, no. 5 (2007): 991-1000.

[166] Chanel, Guillaume, Karim Ansari-Asl, and Thierry Pun. "Valence-arousal evaluation using physiological signals in an emotion recall paradigm." In Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on, pp. 2662-2667. IEEE, 2007.

[167] Petrantonakis, Panagiotis C., and Leontios J. Hadjileontiadis. "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis." Affective Computing, IEEE Transactions on 1, no. 2 (2010): 81-97.

[168] Jeong, Sang-Hoon. "Suggestion of Methods for Understanding User’s Emotional Changes while using a product." In Human Interface and the Management of Information. Methods, Techniques and Tools in Information Design, pp. 59-67. Springer Berlin Heidelberg, 2007.

[169] Blain, Stefanie, Tom Chau, and Alex Mihailidis. "Peripheral autonomic signals as access pathways for individuals with severe disabilities: a literature appraisal." The Open Rehabilitation Journal 1, no. 11 (2008): 27-37.

[170] Poh, Ming-Zher, Nicholas C. Swenson, and Rosalind W. Picard. "A wearable sensor for unobtrusive, long-term assessment of electrodermal activity." Biomedical Engineering, IEEE Transactions on 57, no. 5 (2010): 1243-1252.

[171] Fowles, Don C., Margaret J. Christie, Robert Edelberg, William W. Grings, David T. Lykken, and Peter H. Venables. "Publication recommendations for electrodermal measurements." Psychophysiology 18, no. 3 (1981): 232-239.

[172] Sequeira, Henrique, Pascal Hot, Laetitia Silvert, and Sylvain Delplanque. "Electrical autonomic correlates of emotion." International Journal of Psychophysiology 71, no. 1 (2009): 50-56.

[173] ROTH, WALTON T., MICHAEL E. DAWSON, and DIANE L. FILION. "Publication recommendations for electrodermal measurements." Psychophysiology 49 (2012): 1017-1034.

[174] Cannon, Walter B. Bodily changes in pain, hunger, fear and rage: An Account of Recent Researches Into the Function of Emotional Excitement 1927. READ BOOKS, 2008.

[175] Newton, Grant, John Paul, and E. W. Bovard Jr. "Effect of emotional stress on finger temperature." Psychological Reports 3, no. h (1957): 341-343.

[176] Carpenter, Finley, and Eugene E. Haddan. "Effects of liked and disliked teachers on student behaviour." ERIC, (1966).

[177] Ekman, Paul, Robert W. Levenson, and Wallace V. Friesen. "Autonomic nervous system activity distinguishes among emotions." Science 221, no. 4616 (1983): 1208-1210.

[178] Kistler, Andreas, Charles Mariauzouls, and Klaus von Berlepsch. "Fingertip temperature as an indicator for sympathetic responses." International Journal of Psychophysiology 29, no. 1 (1998): 35-41.

[179] Ward, R. D., P. H. Marsden, Bernadette Cahill, and Clive Johnson. "Physiological responses to well-designed and poorly designed interfaces." In Proceedings of CHI 2002 Workshop on Physiological Computing. 2002.`

[180] Baumgartner, Thomas, Michaela Esslen, and Lutz Jäncke. "From emotion perception to emotion experience: Emotions evoked by pictures and classical music." International Journal of Psychophysiology 60, no. 1 (2006): 34-43.

[181] Nacke, Lennart E., Mark N. Grimshaw, and Craig A. Lindley. "More than a feeling: Measurement of sonic user experience and psychophysiology in a first-person shooter game." Interacting with Computers 22, no. 5 (2010): 336-343.

[182] Mandryk, Regan L., M. Stella Atkins, and Kori M. Inkpen. "A continuous and objective evaluation of emotional experience with interactive play environments." In Proceedings of the SIGCHI conference on Human Factors in computing systems, pp. 1027-1036. ACM, 2006.

[183] Mandryk, Regan L., Kori M. Inkpen, and Thomas W. Calvert. "Using psychophysiological techniques to measure user experience with entertainment technologies." Behaviour & Information Technology 25, no. 2 (2006): 141-158.

[184] Handri, Santoso, Kuniaki Yajima, Shusaku Nomura, Nobuyuki Ogawa, Yoshimasa Kurosawa, and Yoshimi Fukumura. "Evaluation of Student's Physiological Response Towards E-Learning Courses Material by Using GSR Sensor." In Computer and

Information Science (ICIS), 2010 IEEE/ACIS 9th International Conference on, pp. 805-810. IEEE, 2010.

[185] Kätsyri, Jari, Niklas Ravaja, and Mikko Salminen. "Aesthetic images modulate emotional responses to reading news messages on a small screen: A psychophysiological investigation." International Journal of Human-Computer Studies 70, no. 1 (2012): 72-87.

[186] Brennan, Julie, Angele McGrady, Kary Whearty, Denis Lynch, Daniel Rapport, and Paul Schaefer. "Emotional status of third year medical students and their responses to a brief intervention." Annals of Behavioral Science and Medical Education 18, no. 2 (2012): 10-14.

[187] Picard, Rosalind W., and Jennifer Healey. "Affective wearables." Personal Technologies 1, no. 4 (1997): 231-240.

[188] Wikipedia. “Affectiva.” available at http: // en. wikipedia. org/ wiki/ affectiva (Accessed: 20-Sep-2010).

[189] Adobe®. “Captivate® 7” Adobe systems (2011).

[190] Adobe®. “Flash” Adobe Systems (2012).

[191] Reallusion. “CrazyTalk Animator Pro” Reallusion Inc. (2011).

[192] Adobe®. “Photoshop CS6.” Adobe systems (2012).

[193] Sionsoftware. “QuickScore Elite Level II.” Sion Software. Available at http:// www. sionsoft.com/ qse10.html (Accessed: 20-Sep-2010).

[194] Adobe®. “Adobe Premiere.” Adobe system (2011).

[195] Denyer, Morgan. "Medical Statistics at a Glance." Journal of Anatomy 216, no. 4 (2010): 543-543.

[196] Brooke, John. "SUS-A quick and dirty usability scale." Usability evaluation in industry 189 (1996): 194.

[197] MacDonald, Craig M., and Michael E. Atwood. "Changing perspectives on evaluation in HCI: past, present, and future." In CHI'13 Extended Abstracts on Human Factors in Computing Systems, pp. 1969-1978. ACM, 2013.

[198] Bangor, Aaron, Philip Kortum, and James Miller. "Determining what individual SUS scores mean: Adding an adjective rating scale." Journal of usability studies 4, no. 3 (2009): 114-123.

[199] Tabbers, Huib K., Rob L. Martens, and Jeroen JG Merriënboer. "Multimedia instructions and cognitive load theory: Effects of modality and cueing." British Journal of Educational Psychology 74, no. 1 (2004): 71-81.

[200] Cacioppo, John T., Louis G. Tassinary, and Gary G. Berntson, eds. “Handbook of psychophysiology.” Vol. 2. New York: Cambridge University Press, 2007.

[201] Vos, Pieter, Paul De Cock, Vera Munde, Katja Petry, Wim Van Den Noortgate, and Bea Maes. "The tell-tale: What do heart rate; skin temperature and skin conductance reveal

about emotions of people with severe and profound intellectual disabilities?." Research in developmental disabilities 33, no. 4 (2012): 1117-1127.

[202] Maxwell, Scott E., and Harold D. Delaney. “Designing experiments and analyzing data: A model comparison perspective.” Vol. 1. Psychology Press, 2004.

[203] Vivaldi, Antonio. “The Four Seasons and Other Violin Concertos in Full Score: Opus 8, Complete.” Courier Dover Publications, 2013.

[204] Krumhansl, Carol L. "An exploratory study of musical emotions and psychophysiology." Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale 51, no. 4 (1997): 336.

[205] Luck, Geoff, Petri Toiviainen, Jaakko Erkkilä, Olivier Lartillot, Kari Riikkilä, Arto Makelä, Kimmo Pyhäluoto, Heikki Raine, Leila Varkila, and Jukka Värri. "Modelling the relationships between emotional responses to, and musical content of, music therapy improvisations." Psychology of Music (2007).

[206] Tabbers, Huib, Liesbeth Kester, Hans Hummel, and Rob Nadolski. "Interface design for digital courses." Integrated E-learning: Implications for pedagogy, technology and organization (2005): 100-111.

[207] Sankey, Michael, and Mehryar Nooriafshar. "Multiple representations in multimedia and e-learning materials: an issue of literacy." Post Pressed (2005): 149-172.

[208] Fletcher, J. D., and Sigmund Tobias. "The multimedia principle." The Cambridge handbook of multimedia learning 117 (2005): 133.

[209] Mayer, Richard E., and Cheryl I. Johnson. "Revising the redundancy principle in multimedia learning." Journal of Educational Psychology 100, no. 2 (2008): 380.

[210] Hassenzahl, M., M. Burmester, and F. Koller. "AttrakDiff: A questionnaire to measure perceived hedonic and pragmatic quality." In Mensch & Computer, pp. 187-196. 2003.